RECONSTRUCTING IMAGES FROM SPARSE DATA

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Kartik Sankaran Ayyer

August 2014

RECONSTRUCTING IMAGES FROM SPARSE DATA

Kartik Sankaran Ayyer, Ph.D.

Cornell University 2014

In this dissertation, the process of recovering structure from sparse data is discussed. Specifically, studies are undertaken for the case of X-ray imaging with weak signal. The limit of interest is when each image (or data frame) is so sparsely populated with X-ray photons that the structure of the object is not discernable. These techniques will be required to perform single molecule imaging using X-ray free electron lasers and serial microcrystallography experiments at synchrotron sources.

An overview of the problem and the reconstruction algorithm is given in Chapter 1. In Chapters 2, 3 and 4, three different experiments are discussed, each with a different imaging geometry. All three experiments have the same property in that there are many data frames all of whom are very sparsely occupied. In Chapter 2, a shadowgraphy experiment is performed with a randomly rotated mask whose projected shadow on the detector is reconstructed from a large number of images with a signal level as low as 2.5 photons per data frame. In Chapter 3, using a computed tomography (CT) setup with unrecorded orientations, the 3D structure of a plastic figure is reconstructed. And finally, in Chapter 4, a weak beam is used to illuminate a crystal and the sparse diffraction pattern is measured. The successful reconstruction of the 3D intensity distribution shows promise for the possibility of serial microcrystallography [5] at conventional syn-

chrotron sources.

Appendix A contains a more detailed discussion of the reconstruction process for the crystallography experiment in Chapter 4. This record is made to enable other parties to reproduce the results as well as to document the intuition used in choosing various reconstruction parameters.

BIOGRAPHICAL SKETCH

Kartik Ayyer[1] was born in Bharuch, Gujarat, India. He started his undergraduate studies at the Indian Institute of Technology, Delhi in 2005 graduating as a Bachelor of Technology with a major in Engineering Physics in 2009. He was then admitted to the PhD program in the Physics department at Cornell University. He started working with his adviser Veit Elser in the summer of 2010 and earned his PhD in 2014.

[1]He prefers not to use his middle name. He only has one due to an overzealous passport office employee who inserted Kartik's father's name without asking.

To my family

ACKNOWLEDGEMENTS

I would like to first thank my adviser, Veit Elser for always keeping an open door for my questions and discussions. I am also very grateful for his emphasis on an understanding of fundamentals and to question assumptions. This has proven especially useful in the field of reconstruction algorithms where new ideas are common, but good ones are not. I have also been enriched by my discussions with Profs. Sol Gruner and Chris Myers, who served on my committee. Sol has also done me a great favor by pushing me to give talks and exposing me to the wider world of crystallographers and their way of thinking.

I also want to acknowledge my senior Elser group members, Duane Loh, Yoav Kallus and Diarmuid Cahalane for providing very helpful career guidance and for showing me the way to think about problems in a more systematic way. My contemporaries Zhen Wah Tan, Hyung Joo Park, Yi Jiang and Ti Yen Lan have stimulated many interesting discussions. Outside of the group, Hugh Philipp, Mark Tate and Jeney Wierman have not only been fantastic collaborators, but have been instrumental in my attempts to be a more rounded physicist with a better intuition of experimental details in X-ray. Also, if I ever try my hand at an experiment, it will be because of the positive impression they gave me of that domain.

My five years in Ithaca have gone by much more quickly that expected. A good part of that has been in the excellent company of Shivam Ghosh, Mihir Khadilkar, Jesse Silverberg, Ines Firmo, Kyung Min Lee, Thomas Bachlechner and Ishita Mukhopadhyay. I am grateful to Sivaram Kalidas, Sampath and Nalini Aiyar and Vivek Iyer for giving open invitations to visit them when I wanted to get

out of the town.

I would like to thank my parents, Jaya and Sankaran Ayyer and my brother Arvind for inculcating a curious nature in me. I was never left out of technical conversations they were having even though I was the youngest by 8 years.

And finally, there is no way I would have survived the last five years without the constant love and support from my wife Avni.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 A gedanken experiment

Imagine an experiment where we want to measure the spatial distribution of a weak X-ray signal from some source. Say there is an object of interest which modulates the intensity, and figuring out the spatial distribution will lead to the determination of that object's structure. X-rays are electromagnetic waves which come in wave packets known as photons, of which we can only count an integer number. So the first thing to do is to grab an X-ray detector; something that can detect when an X-ray photon hits it. Now we stick this detector in the path of the beam. We also need to set a time interval, t_0, in which we want to make a measurement. This is necessary because if we wait infinitely long we will get infinite photons no matter what. We define a weak X-ray signal as that in which the average number of photons in that time interval (known as the fluence) is much less than one. That means that most of the time, we get nothing. Once in a while we will get one photon, and even more rarely, we will get two or more.

Now imagine a detector with a million such little detectors next to each other (call them pixels). Since there are so many pixels, in each time interval, we do get a few photons and the distribution of photons is spatially related to the distribution of intensities i.e. the higher the intensity at a given point, the more likely it

is that a photon was captured. Thus, given enough measurements, one should be able to determine the spatial distribution of intensities, solving the problem. Note that there isn't enough signal in one image to determine the distribution, but as a whole, one has enough information. Of course, this is an oversimplification because there could be other sources of noise and background, one being the detector itself. Another is that the intensity modulations may not all be from the object of interest, but from something else in the path of the beam (background).

However, the experiment is much harder than just reducing the amount of measurement noise and background. Someone tells you that the object of interest itself is different about from image to image. That means that if you just average all the data frames (the images), you are only reconstructing the average distribution, and that is not good enough. One common example of such missing information is that the object has an unknown orientation. Each time you are looking at the object from a different viewpoint, meaning that the intensity modulation caused by it is different. You are effectively being asked to reconstruct many different objects using one unsorted data set. Thus, to solve the problem you need to identify which frames come from which orientations and only average like-oriented frames.

The naive approach would be to look for data frames which are similar and say that they must be from similar orientations. This has been explored [6, 8], and in some cases it will work. These methods rely on cross correlations between different frames and creating a graph whose edges correspond to pairs of frames

with high correlation. However, for the signal levels we have in our experiment ($\sim 10^{-3}$ photons/pixel/frame), the cross-correlations will mostly be zero, and in the few rare occasions where they are non-zero, it is more likely to be a result of coincidence than due to similar orientations.

1.2 Real world applications

Before moving on to discuss our approach and experiments that have been performed to demonstrate its effectiveness, let's take a look at examples of where these situations are encountered. X-rays have been used in what is called diffractive imaging for many decades. One major application is crystallography, where crystals of complex molecules like proteins are illuminated by a strong X-ray beam. These crystals behave like 3D versions of the double slit experiment and generate a complex interference pattern which can be analyzed to determine the structure of the molecules. This method has been very successful and has been of exceptional value to biology. However, the hardest step in these experiments is the creation of a large enough crystal that will give good diffraction patterns. The amount of diffracted intensity generated by an object loosely depends on the strength of the beam and the size of the particle. Of course the recorded signal also depends on how long the crystal is exposed to the beam. Unfortunately, there is also a limit on this due to radiation damage. This means that as X-rays are incident on a material, they start to affect the molecules and change their structure in

unpredictable and non-uniform ways. The result of all of this is that it has been estimated [10] that with conventional storage ring-based X-ray sources, the smallest a crystal can be to give a full data set is $2\mu m^3$. This is the best case and requires that the crystal be cooled to around 100K, which is not ideal in terms of biology.

One solution to this problem is to build a brighter source. This has been done with X-ray free electron lasers (XFELs) and smaller crystals have been imaged using snapshot images [5, 4]. This has been termed serial crystallography. Here, each data frame has enough signal that it is easy to identify a Bragg pattern. There is enough information in these patterns to allow algorithms to assign a Miller index (hkl) to each spot, and from there to figure out the orientation [28]. If this experiment is repeated at what is known as a 3rd generation synchrotron source, we get the kind of sparse data outlined above. Another avenue to explore would be to use XFELs to image single molecules, again leading to sparse data and unknown orientations. These applications would allow the solving of structures for proteins which are not currently solvable.

1.3 Orienting sparse data frames

The approach outlined below was first demonstrated in [15, 7] and was given the name EMC algorithm. This is an iterative Bayesian algorithm applying the principle of expectation-maximization [2] to collectively assign orientations to all data frames and simultaneously reconstruct the common spatial distribution which the

various orientations are sampling. EMC stands for Expand-Maximize-Compress, which are the three steps in each iteration. The principal idea is compare the current model of the intensity distribution, for various views, with the data frames and assign probabilities to them. Using these probabilities, one updates the views by maximizing the likelihood of them generating the data frames. The Expand and Compress steps convert from and to the common spatial intensity distribution, which may be three-dimensional, to the views which contain the mean photon counts per pixel. The exact process by which this conversion takes place depends upon the geometry of the experimental setup as well as the nature of the missing information, be it orientation, translation or some other variable.

The Maximize step is where these views are compared with the data and updated. This step can itself be broken down into two sub-steps, expectation and maximization. Before that, a quick note about shot noise. If the average intensity at a pixel is given by w, the probability of capturing k photons is given by the Poisson distribution

$$P(k) = \frac{w^k e^{-w}}{k!} \tag{1.1}$$

Let there be R discrete views represented by the index r and D data frames represented by the index d. In the expectation sub-step, one calculates P_{dr}, the probability of frame d being generated by the object in orientation r. If the intensity at the pixel t in orientation r is given by W_{rt} and the number of photons in pixel t of frame d is given by K_{dt}

$$P_{dr} = \frac{\ell_{dr}}{\sum_r \ell_{dr}} \tag{1.2}$$

where ℓ_{dr} is the likelihood given by

$$\ell_{dr} = \prod_t W_{rt}^{K_{dt}} e^{-W_{rt}} \tag{1.3}$$

This is just a reformulation of Equation 1.1 with the additional assumption that the Poisson process at each pixel is independent. The factorial part is ignored as it cancels out on normalization. The normalization is done to ensure that the probability of the frame d having any orientation is 1. Once these probabilities are calculated using the current model, the updated views, W'_{rt} are generated by maximizing the likelihood of generating the data. The log-likelihood function maximized is

$$\log[Q(W'_{rt})] = \sum_d \sum_r \sum_t [P_{dr}(K_{dt} \log(W'_{rt}) - W'_{rt})] \tag{1.4}$$

On maximizing with respect to W'_{rt}, we obtain the intuitive update rule

$$W'_{rt} = \frac{\sum_d P_{dr} K_{dt}}{\sum_d P_{dr}} \tag{1.5}$$

As can be readily seen, this is just the weighted mean of the data frames with the weights being the probabilities calculated using the current model.

One starts with a random initial guess for the model and then the EMC steps are applied to update the model till it stops changing. Once this happens, the model is said to have converged and some sanity checks are applied to confirm that the final model is the true solution. The checks depend on the particular experiment and will be discussed in more detail later.

1.4 Proof-of-principle experiments

The rest of this thesis describes three experiments, with three different geometries, all of which share the features of sparse data frames and unknown orientation.

The first is shadowgraphy, where a lead mask was rotated in-plane in the path of a very weak x-ray beam. With signal levels as low as 2.5 photons/frame, the shadow of the mask is reconstructed. This experiment was the easiest to analyze, both technically as well as in terms of the requirement of photons/frame. The effect of additional background was also analyzed by adding numerically generated background photons with a uniform distribution to each data frame.

The second experiment was real-space tomography, where a plastic figure was illuminated, again by a weak beam. In this case the intensity modulation was caused by attenuation inside the object. This was a much lower contrast experiment, in addition to being technically challenging due to the non-linear nature of the attenuation process. There were special intricacies in executing the Expand and Compress steps. Another feature observed was the requirement of a minimum number of photons/frame below which no orientations could be assigned.

The third experiment was crystallography, where a rotating crystal was used to generate sparse diffraction patterns. This is a relevant experiment as it points to the possibility of performing serial crystallography experiments with synchrotron sources. Since background scattering is of crucial importance in crystallography experiments, this aspect was also studied here. On the technical side, it was found

that some prior information was required about the crystal (the unit cell parameters) in order to converge to the right solution.

In the appendix, there is a detailed discussion of the reconstruction process for the crystallography experiment and the reasoning behind some of the choices made. This chapter was prepared to allow others to reproduce the results in detail, rather than just the general structure of the computations. Another motivation is to emphasize the point that in the current age of non-trivial data analysis, there should be a discussion of the process, and not just the final polished result, just as in good experimental and theoretical discussions.

CHAPTER 2

SHADOWGRAPHY WITH SPARSE DATA

The contents of this chapter have been published in *Optics Express* with coauthors Hugh Philipp, Mark Tate, Veit Elser and Sol Gruner [19].

2.1 Abstract

Single-particle imaging experiments of biomolecules at x-ray free-electron lasers (XFELs) require processing hundreds of thousands of images that contain very few x-rays. Each low-fluence image of the diffraction pattern is produced by a single, randomly oriented particle, such as a protein. We demonstrate the feasibility of recovering structural information at these extremes using low-fluence images of a randomly oriented 2D x-ray mask. Successful reconstruction is obtained with images averaging only 2.5 photons per frame, where it seems doubtful there could be information about the state of rotation, let alone the image contrast. This is accomplished with an expectation maximization algorithm that processes the low-fluence data in aggregate, and without any prior knowledge of the object or its orientation. The versatility of the method promises, more generally, to redefine what measurement scenarios can provide useful signal.

2.2 Introduction

Ultra-intense, ultra-fast x-ray pulses from x-ray free electron lasers (XFELs), such as the Linac Coherent Light Source (LCLS), hold potential to provide structural information about proteins for which crystals are unavailable [16]. This so-called single particle imaging experiment involves scattering x-rays off many copies of a given protein, one protein at a time. This yields a sequence of x-ray detector images, each resulting from the diffraction of a single pulse scattered off a single protein. Since only some thousands of x-rays are scattered off each protein, each x-ray image, which consists of a few million pixels, is very sparsely populated with data: On average, each pixel in each image receives far fewer than one x-ray. If each protein intersected the x-ray beam in the same orientation, one could simply add many images to recover a statistically significant data set. However, each protein intersects the x-ray pulses in unknown, random orientations and there are too few x-rays in any single image to determine the orientation of that protein. The challenge is to devise a method that combines information from a large number of these severely Poisson noise limited images into a complete, consistently oriented data set.

A data reduction method has been proposed [15, 14] that should work, in principle, even in the case where the number of x-rays per image is so small that it is impossible to discover similar orientations by cross-correlating images. However, this method has not been experimentally tested in the extreme case of only a few photons per image. Here, we demonstrate that a simplified 2D version of the algo-

rithm is capable of reducing this kind of data even with images that average only 2.5 x-ray photons per frame ($\sim 10^{-4}$ photons per pixel per frame). For this demonstration, in order to emulate realistic detector noise performance, we use the same pixel array detector chip that makes up the detector installed at the LCLS Coherent X-ray Imaging (CXI) beamline for the protein imaging experiment [18, 20]. The CXI detector differs from the one used here in that the former is an array of these chips to cover a larger area.

Our experiment simulates the diffraction patterns of randomly rotated particles by means of a mask placed in front of the detector that is given random rotations and is uniformly illuminated by a highly attenuated beam. We argue that this experiment contains all the salient features of a full 3D intensity reconstruction, namely unknown sample orientations and very low signal. The additional features of the single particle imaging experiment are the extra degrees of rotational freedom of the sample in 3D and the increased size of the detector. Since we discretize the space of rotations, the additional degrees of freedom only increases the size of the discrete set of orientations required to adequately sample the space. This does not make a qualitative difference to the problem. Similarly, a larger detector only increases the computational resources required. The reconstruction algorithm scales linearly as the number of orientational samples and the total number of photons imaged. Simulations have shown that the higher computational requirements can be handled comfortably [15].

2.3 Expectation maximization algorithm

The algorithm for reconstructing the x-ray intensity from data follows the expectation maximization (EM) principle [2]. EM starts with a random model of the intensity $w(i)$, where each pixel i is assigned a random value uniformly in the range $[0, 1]$. These values are iteratively updated by a rule that can only increase the likelihood of the model. The initial model is random because no information about the model is known.

Each iteration involves two steps. In the first step, each frame of data, f, is assigned a probability distribution, $p_f(r)$, with respect to its unknown rotation, r, relative to the current intensity model. The rotations are sampled in increments of $2\pi/N$, where N defines the angular resolution of the reconstruction.

Each frame comprises photon occupancy, $k_f(i)$, at pixel i, which in our low-fluence experiment are almost all zero, the exceptions being equal to 1. Because the photon counts are independent Poisson samples of the intensity at each pixel, the probability is

$$p_f(r) \propto \prod_i \frac{w(i+r)^{k_f(i)}}{k_f(i)!} e^{-w(i+r)} \propto \prod_{i \in I_f} w(i+r), \tag{2.1}$$

where $i + r$ is rotation r, applied to pixel i, I_f is the set of pixels recording photons in frame f, with the final expression applying in the low-fluence limit. After normalizing the distributions, $p_f(r)$, the algorithm proceeds to the second step.

In the second step the algorithm aggregates the photon data from all the frames according to the distributions obtained in the first step:

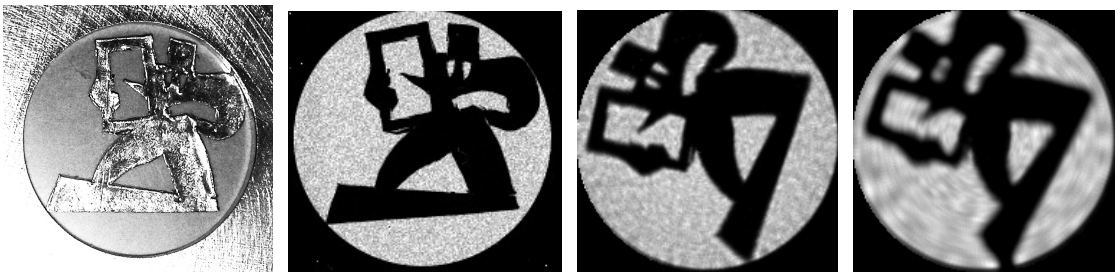
$$w'(i) = \left\langle \sum_r p_f(r) k_f(i-r) \right\rangle_f . \tag{2.2}$$

The updated intensity model $w'(i)$ is an average of the photon counts in all frames with the appropriate distribution of rotations applied to each one. Linear interpolation is used in both steps, when mapping one square grid (model) onto one that has been rotated (detector). We consider the reconstruction to have converged when the root-mean-square difference between successive models is below a cutoff. The EM algorithm is still valid when the data is winnowed by a structure-neutral criterion, such as the photon occupancy. In our implementation, for example, we discarded all frames with zero occupancy.

An alternative approach being considered[11] for determining the rotations of randomly oriented particles involves classifying data on the basis of cross-correlations:

$$c_{ff'} = \sum_i k_f(i) k_{f'}(i). \tag{2.3}$$

This is not viable in the ultra-low fluence limit because the numbers $c_{ff'}$ are essentially all zero, and in any event do not distinguish frames derived from like or unlike particle orientations. A proposal [6, 8] to orient data by means of diffusion nodes on a graph associated with the data is not suitable for the low-fluence case we are considering.



(a) (b) (c) (d)

Figure 2.1. (a) The lead x-ray mask mounted within an aperture in an aluminum disk. (b) A static x-ray image of the pattern collected as 432 individual frames with approximately 1/5 photon per pixel per frame. The frames were thresholded and averaged. (c) A reconstruction using randomly-oriented data having an average 11.5 photons/frame and 1.2 million recorded photons. (d) A reconstruction using randomly-oriented data having an average 2.5 photons/frame and 1.2 million recorded photons.

For our experiment such a graph comprises nearly a million nodes (one per frame) but only a handful of edges (rare instances where $c_{ff'} > 0$). The EM algorithm, by contrast, compares each frame not with other frames but with a model. A greater sensitivity of rotation determination in the EM algorithm can be traced to the multiplicative nature of the comparison expressed by Eq. 2.1.

The EM algorithm should in principle work with data of arbitrarily low-fluence. It is clear that this is the case when we consider that there will be rare fluctuations where the photon occupancy is 2 or greater, even when the mean is just a fraction of a photon. A fair assessment of the viability of reconstructions in the low-fluence regime must therefore take into account the inevitability of background. The effects of background in the interpretation of our results are well

captured by a simple information rate ratio:

$$R = \frac{\sigma(1 + \mathrm{SN}^{-1}) \log(1 + \mathrm{SN}) - (\sigma + \mathrm{SN}^{-1}) \log(1 + \sigma\,\mathrm{SN})}{-\sigma \log \sigma}. \tag{2.4}$$

This is the information rate at signal-to-noise SN (the ratio of signal to background fluence) divided by the rate in the absence of background, where σ is the fraction of uncovered pixels.

Eq. 2.4 is based on prior distributions for both the signal and background; a derivation is given in the appendix. The signal for our experiment was modeled as two-valued, corresponding to zero for pixels covered by the mask and a constant nonzero value for uncovered pixels. A single parameter, the fraction σ of uncovered pixels, describes this signal prior. When modeling a true diffraction signal, the prior would instead be the Wilson distribution. To model the background, we used a single-valued distribution corresponding to uniform background counts across the detector.

As an example of the application of Eq. 2.4, consider the case of $\mathrm{SN} = 1/10$, which for $\sigma = 0.6$ gives $R \approx 0.01$. A low fluence experiment with 2 signal photons per frame and this poor signal-to-noise would therefore be like a zero-background experiment with only $2R = 0.02$ photons per frame. Applying Poisson statistics to this low rate we find that only about 1 in 5000 frames would have 2 or more photons and be actually useful for the reconstruction.

2.4 Data collection

To test the reconstruction algorithm with experimental data from randomly oriented samples, a pattern was cut out of x-ray opaque lead sheet to create an x-ray shadow mask (Fig. 2.1(a)). This mask was then mounted within a 19 mm diameter aperture of an opaque metal disk that fit onto a rotation stage (Newport URS100BPP) with the center of rotation approximately at the center of the aperture.

A very low-power copper anode x-ray tube was used to generate x-rays (TruFocus TFS 6050 Cu, 50 W maximum). It was operated at an anode voltage of 10.1 kV to reduce high-energy bremsstrahlung. A 50 micron thick nickel filter was used to preferentially remove the K_β of the tube spectrum to produce an approximately monochromatic x-ray beam of 8-keV Cu K_α radiation. The rotation stage and aperture were mounted on the end of a 45 cm flight-path to produce a nearly flat-field illumination of x-rays across the 19 mm sample.

The x-ray mask and a static x-ray image of the pattern are shown in Fig. 2.1(a) and Fig. 2.1(b). Cornell's LCLS Pixel Array Detector (PAD), comprising a single-chip 194 x 185 pixel array, was placed after the mask along the flight path. The PAD was positioned so the entirety of the aperture image was incident on the detector. The x-ray image shown in Fig. 2.1(b) was collected by summing 432 images of the mask at fixed position. Each 0.1 s image had an occupancy of approximately 0.2 x-ray photons per pixel per frame in the unobstructed regions.

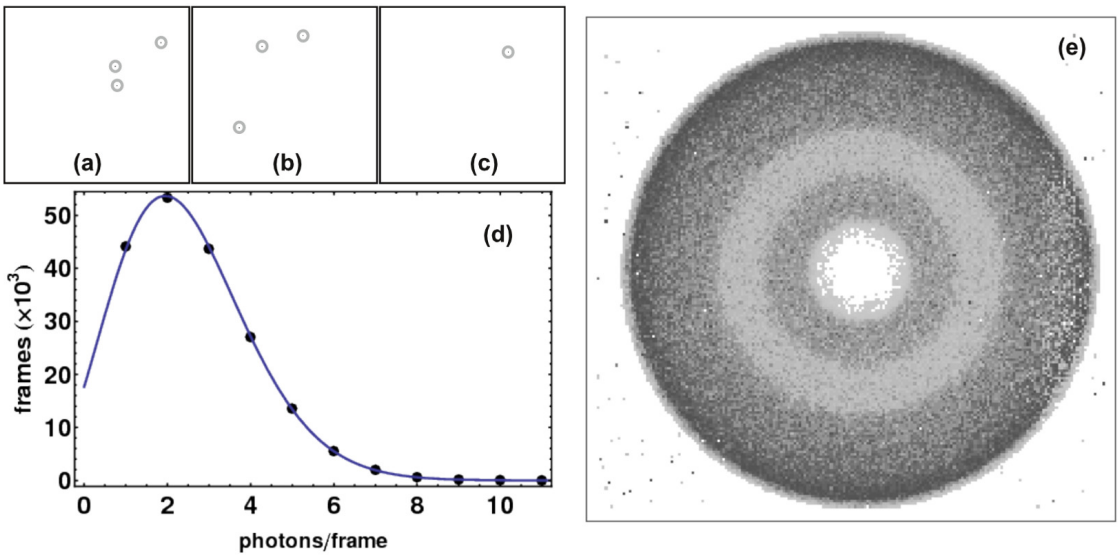


Figure 2.2. (a-c) Three sample frames from the 2.5 photon/frame data set with detected x-ray photons circled. (d) Occupancy histogram compared with the Poisson distribution. (e) The sum of all thresholded frames from the 2.5 photon/frame data set showing a uniform angular distribution of data.

Very low-fluence data were also collected with a continuously rotating sample. The detector collected images at 100 frames per second with a per-frame integration time of 100 microseconds. The waveforms used for digitization and detector readout were the same as those used when the detector is running at 120 frames per second, the frame rate of the LCLS, except the internal trigger was set to a 10 ms period (rather than 8 ms). X-ray tube currents of 0.15 mA, and 0.03 mA were used to produce varying x-ray intensities. With each current, hundreds of thousands of images were collected. Figures 2(a-c) show three typical very low-fluence mask images, in each case consisting of only a few x-rays per frame.

Dark signal measurements were made throughout the data collection sequence

by periodically taking groups of 144 frames with the x-ray shutter closed. These dark frames were used to define a low-noise zero-level which was subtracted from individual signal frames to extract the x-ray induced signal from the raw detector output.

Analog integrating detectors are required at XFELs because many experiments deliver more than one x-ray photon per pixel per frame (as expected at low scattering angles in single particle imaging experiments), and the x-ray pulse is too short for photon counting electronics. Minimum signal threshold values can be applied to the analog data to reject low-level noise [20]. The threshold in this experiment was set to 0.7 x-ray photons (for 2.5 photon/frame data set) or 0.75 x-ray photons (for the 11.5 photon/frame data set). At these thresholds approximately 0.05 and 0.01 false positive photon measurements per frame are expected, using a normal distribution and the previously measured [18] pixel signal-to-noise ratio of 7 for a single 8-keV x-ray. The lower threshold was used for 2.5 photon/frame data because this was the last data set taken and progressively less favorable parameters were chosen to test the robustness of the detector and algorithm. No compensation was used for charge sharing between adjacent pixels, nor were pixel gains individually calibrated. A single, global threshold and nominal pixel gain value were applied across the array.

Separate data sets analyzed included hundreds of thousands of frames with mean fluences of 11.5 photons/frame and 2.5 photons/frame.

2.5 Results

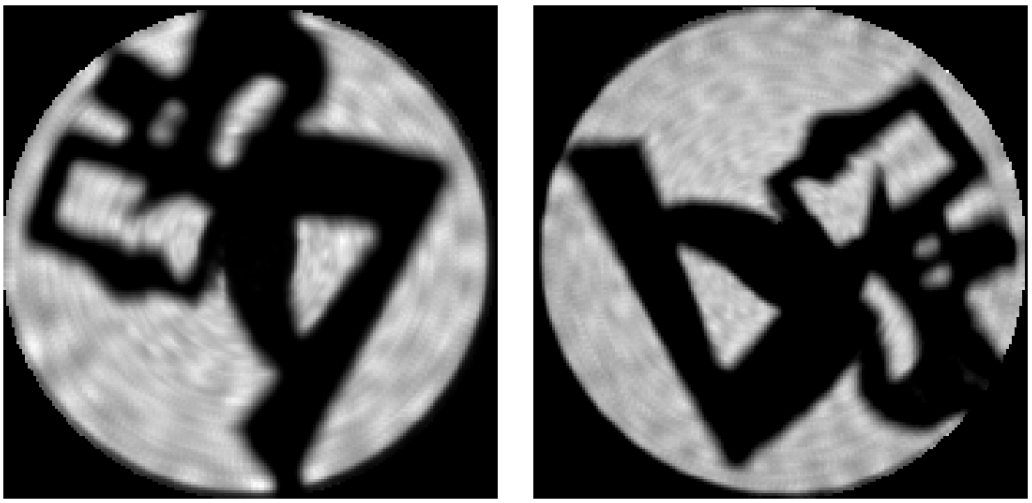
Reconstructed images are shown in Fig. 2.1(c) and Fig. 2.1(d). Figure 2.1(d) was reconstructed using 450,000 frames of data with an average of 2.5 photons per frame. The reconstruction algorithm used 180 equally spaced 2° steps. Figure 2.2(e) shows a simple sum of the thresholded frames of data that results in a rotationally smeared image with a uniform angular distribution. Figure 2.2(d) shows a per-frame occupancy histogram, confirming the expected Poisson distribution. This data set has a total of 1.2 million photons. For comparison, a data set with a similar total number of photons, but a higher per-frame photon average (and thus fewer frames) was also processed. The reconstruction is shown in Fig. 2.1(c), where the average occupancy was 11.5 photons/frame.

The quality of the two reconstructions differs in both spatial resolution and contrast, with the 11.5 photons/frame data yielding better results. This agrees with the results of reconstructing 3D intensities from simulated single-particle diffraction data [15], also at very low fluence. The degradation in quality occurs when a significant fraction of the information content in each frame, about half, is just the orientational state. There is a sharp increase in the iteration count of the EM algorithm when this criterion is met: the 2.5 photon/frame data required 220 iterations, compared to 49 iterations for the 11.5 photon/frame data. The convergence of these 2D reconstructions is similar to the performance of the algorithm in 3D with simulated data in the ultra-low-fluence limit. This provides further confirmation that our test scenario is directly relevant to single protein imaging

at an XFEL. Supplementary materials associated with this paper include a movie showing progression of the model through 300 iterations from initial guess to converged solution.[1]

By adding a uniform distribution of computer generated photon counts to the data sets, and processing it by the EM algorithm as before, we are able to simulate the effects of SN = 1 background scattering from gas molecules along the path of the incident x-ray beam in single particle experiments. This should be the major source of background signal and many times larger than the detector noise when the detector data are properly thresholded. Not surprisingly we find deterioration in the quality of the reconstruction. The degree of degradation is consistent with the information ratio R quoted above, which equals 0.26 for our chosen signal-to-noise. With this level of background our data set with 11.5 signal-photons/frame corresponds to a zero-background data set with only 3 photons/frame. The resulting reconstruction by the EM algorithm was therefore similar to that of our 2.5 photons/frame background-free reconstruction in both image quality and number of iterations (Fig. 2.3).

[1]This movie can be accessed from the webpage for [19] in the multimedia section at http://www.opticsinfobase.org/oe/fulltext.cfm?uri=oe-20-12-13129&id=234556



(a) (b)

Figure 2.3. Effect of background on reconstruction quality. (a) Reconstruction from 2.5 photons/frame data set and no added background. This is the same as Fig. 2.1(d). (b) Reconstruction from the 11.5 photons/frame data set with an average of 11.5 photons of background added per frame 'by hand' with a Poisson distribution. The background level was subtracted before rendering to facilitate comparison to (a). As can be seen, the quality of the reconstructions is about the same, and much reduced from the original 11.5 photons/frame data (Fig. 2.1(c)).

2.6 Conclusion

Although this demonstration was motivated by the ongoing effort to realize single particle imaging, the strategy we employed applies more generally to measurements which seek to eliminate ensemble averaging and as a result yield extremely weak signals. Temporal averaging is avoided by short pulses of illumination and the spatial counterpart is achieved by isolation (e.g. single particles) or focusing, as in the case of ptychography [25]. In all these cases one sacrifices signal within a single frame, thus putting an increased burden on the recording of weak signals

with high fidelity and reconstructing from the resulting very sparse data. The envisioned single particle experiments at XFELs are an extreme example of this, but the same approach would also apply to experiments with lower intensity sources, for example, Energy Recovery Linac (ERL) x-ray sources [3]. An ERL can deliver very short x-ray pulses that are much less intense than XFEL pulses, but deliver many more pulses per second to compensate. Ptychography performed with an ERL, in conjunction with our data acquisition/analysis method, looks especially promising. Data acquisition would be fast and yet immune to mechanical instabilities because of the short pulse duration, while jitter in the position of the focus would be algorithmically reconstructed, in analogy with the angular reconstructions in our demonstration.

Appendix: Information reduction by background

The effect of background in low-flux x-ray measurements is well modeled by a noisy communications channel [22]. Each detector pixel represents one channel and the noise analysis can be carried out for a single channel because the noise processes are, to a good approximation, already independent at the level of pixels[2].

Let w be the fluence of radiation integrated over one pixel in the time interval of a single frame. In our experiment w takes two values: the background value v

[2]Background scattering from gas molecules is incoherent and uncorrelated between pixels, as is the photo-absorption process that in effect samples the number of photons in the radiation field.

when the pixel is covered by mask, and $\nu+\mu$ when the pixel is uncovered. Because the mask is given random rotations, the prior distribution on w is,

$$p(w) = (1-\sigma)\delta(w-\nu) + \sigma\delta(w-\nu-\mu), \tag{2.5}$$

where σ represents the fraction of open area in the mask.

The detector pixel measures w as a discrete number of photons k. This is a Poisson process with conditional probability

$$p(k|w) = \frac{w^k}{k!}e^{-w}. \tag{2.6}$$

In the communications analogy k is the message that is received when w is sent. The information obtained in the measurement (transmitted by the channel) equals the mutual information I associated with the joint probability distribution

$$p(w,k) = p(w)p(k|w). \tag{2.7}$$

The mutual information is the difference of entropies

$$I = H_k - H_{k|w}, \tag{2.8}$$

where

$$H_k = \sum_k -p(k)\log p(k) \tag{2.9}$$

is the entropy of the photon counts and

$$H_{k|w} = \int dw\, p(w) \sum_k -p(k|w)\log p(k|w) \tag{2.10}$$

is the average entropy of the counts when the flux is given.

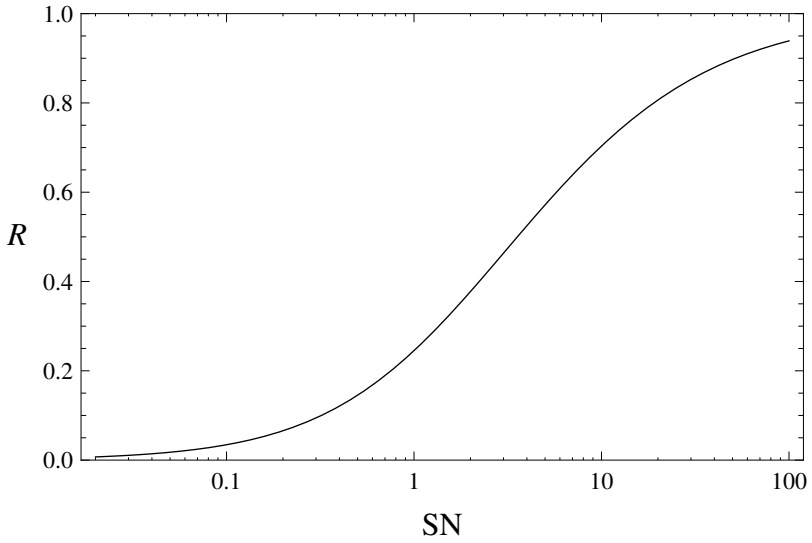


Figure 2.4. Reduction in the rate at which pixels measuring photons acquire information as a function of signal-to-noise. Plotted is the ratio R of the information rate, acquired with and without background. SN is the ratio of signal to background photon counts. This plot applies to the model where half the pixels receive only background (covered by mask) and the other half receive signal and background (not covered by mask).

Both entropies simplify considerably in the extreme low flux limit ($\nu \to 0$, $\mu \to 0$) where we can neglect $k > 1$ in the sums:

$$I(\mu, \nu) = \mu \left(\sigma(1 + \mathrm{SN}^{-1}) \log(1 + \mathrm{SN}) - (\sigma + \mathrm{SN}^{-1}) \log(1 + \sigma\,\mathrm{SN}) \right) + O(\mu^2). \qquad (2.11)$$

Here $\mathrm{SN} = \mu/\nu$ is the signal-to-noise. The zero background limit at low signal flux corresponds to the limit of infinite SN in the expression above:

$$I(\mu, 0) = -\mu\,\sigma \log \sigma + O(\mu^2). \qquad (2.12)$$

The ratio of these quantities depends only on SN and represents the ratio of the

information rate is acquired with and without background:

$$\frac{I(\mu,\nu)}{I(\mu,0)} = R(\mathrm{SN}) = \frac{\sigma(1+\mathrm{SN}^{-1})\log(1+\mathrm{SN}) - (\sigma+\mathrm{SN}^{-1})\log(1+\sigma\,\mathrm{SN})}{-\sigma\log\sigma}. \qquad (2.13)$$

The function R is linear at small SN and monotonically approaches 1 at large SN. Figure 4 shows a plot for the case $\sigma = 1/2$.

Acknowledgments

LCLS PAD development was supported by subcontract from SLAC under DOE Contract DE-AC02-76SF00515. Detector development at Cornell is also supported by DOE Grants FG02-97ER62443, DE-FG02-10ER46693 and the Keck Foundation. CHESS is supported by NSF and NIH-NIGMS under NSF Grant DMR-0936384. The data analysis dealing with the extraction from randomly-oriented, sparse data is supported by DOE Grant DE-FG02-11ER16210.

CHAPTER 3

TOMOGRAPHY WITH SPARSE DATA

The contents of this chapter have been published in *Optics Express* with coauthors Hugh Philipp, Mark Tate, Veit Elser and Sol Gruner [1].

3.1 Abstract

Schemes for X-ray imaging single protein molecules using new x-ray sources, like x-ray free electron lasers (XFELs), require processing many frames of data that are obtained by taking temporally short snapshots of identical molecules, each with a random and unknown orientation. Due to the small size of the molecules and short exposure times, average signal levels of much less than 1 photon/pixel/frame are expected, much too low to be processed using standard methods. One approach to process the data is to use statistical methods developed in the EMC algorithm [15] which processes the data set as a whole. In this paper we apply this method to a real-space tomographic reconstruction using sparse frames of data (below 10^{-2} photons/pixel/frame) obtained by performing x-ray transmission measurements of a low-contrast, randomly-oriented object. This extends the work by Philipp et al. [19] to three dimensions and is one step closer to the single molecule reconstruction problem.

3.2 Introduction

X-ray free electron lasers (XFELs) have been successful in performing crystallographic reconstructions of biomolecules with nanocrystalline samples having as few as 10^3 unit cells[5, 4]. This is possible, in part, because the high signal and Bragg peak concentration allows indexing methods to determine the orientation of each frame [28]. There are, however, many situations in which indexing with a single frame is not possible either because of the nature of the sample (e.g. a non-crystalline particle or protein) or because the number of scattered photons detected in a single frame simply do not provide enough information. In these cases, a different approach is required.

The EMC (expand-maximize-compress) algorithm[15, 14] presents a method of dealing with these more difficult data sets. The heart of this algorithm depends on expectation maximization (EM) methods that were experimentally demonstrated previously[19] using the prototype detector based upon the ASIC (application specific integrated circuit) developed by Cornell for the CS-PAD detector at the Linac Coherent Light Source (LCLS).

The present work extends this approach to 3D tomographic reconstructions using sparse x-ray transmission data collected from a 5 cm sized object, where each data frame is from a random and unknown orientation. The sparse data frames used for reconstruction have signal levels of $\sim 10^{-3} - 10^{-2}$ photons/pixel/frame. The detector system used is a tiled pixel array detector (PAD). Each tile has

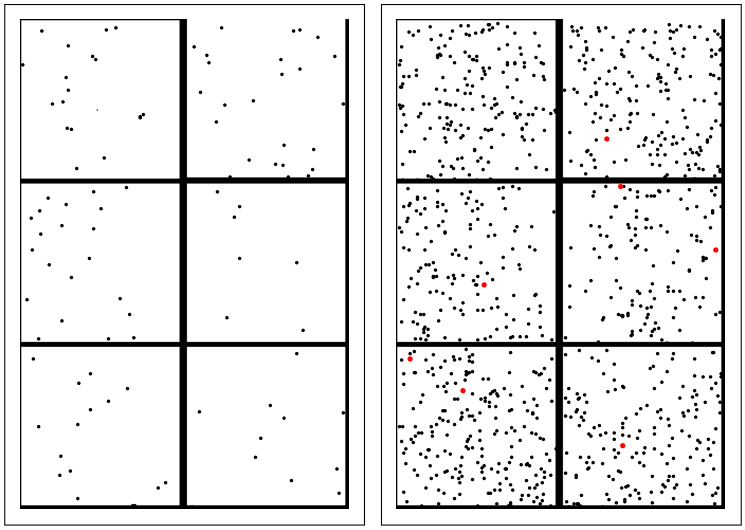
128×128 pixels and uses the mixed-mode pixel array detector (MMPAD) ASIC[26]. The overall tiling format is a 3×2 grid[24]. The object reconstructed is relatively low-contrast at the K_α emission line of molybdenum (17.5 keV). The object absorbs from 5% of the photons incident on it in the thinnest regions to 90% in the central portion.

3.3 Data generation

Low-signal tomographic reconstructions using the EMC algorithm were tested using a plastic figure, roughly matching the size of the detector, as the object of study. A 50 watt molybdenum anode x-ray tube was used to generate x-rays (Tru-Focus TCM-5000M). The applied high voltage was 21.5 kV and the current was set to 0.05 mA. A 400 micron zirconium filter was used to better isolate the K line of molybdenum (17.5 keV) from the x-ray tube emission spectrum by preferentially attenuating lower energies and higher energies (beyond the K-edge of zirconium, 18.0 keV). The distance between the x-ray source and the sample was 1.3 meters. The detector was placed directly after the sample. Detector pixels measured 150 μm × 150 μm. The per-frame exposure time was chosen to be 4 ms with an average signal of 100 photons per frame (10^{-3} photons/pixel/frame).

The plastic figure (Fig. 3.1(a)) was mounted on a post and attached to a rotation stage (Newport URS100BPP). During data taking, the rotation stage turned continuously at a rate of 2 degrees per second. The data was taken continuously

(a)



(b) (c)

Figure 3.1. (a) Photographs of the target object, a plastic toy figure about 50 mm tall. (b) Typical frame of data containing 96 photons in a 396x266 pixel detector. This translates to 9.1×10^{-4} photons/pixel. (c) Data frame with 1025 photons obtained by combining 10 consecutive frames from the previous data set. The sizes of the pixels recording photons have been enlarged to improve visibility. Pixels with two photons are shown in red.

with 860 μs between frames in batches of 1000 frames. Variable time delays were used between batches, to ensure that there was no time sequence bias, and 15.6 million frames were acquired.

Each frame (Fig. 3.1(b)) acquired was converted to photon counts in the following way: 1) An average of 100 to 300 dark frames was subtracted from the signal frames and detector specific offsets were corrected. 2) A threshold was applied to each pixel that corresponded to 60% of a single molybdenum K_α x-ray. Pixels that did not exceed this threshold were set to zero signal level. 3) The number of photons detected in a pixel exceeding the threshold was determined by dividing the pixel signal by the single photon signal level, and rounding to the nearest integer. Note the edge pixels around the rim of each tile are larger than the interior pixels due to edge effects in the sensor[9]. No correction for this effect was applied to the data.

After digitizing each frame, a list of pixels having non-zero photons was recorded. Because of the sparse nature of the low-fluence data, recording pixel coordinate and number of photons, rather than all pixel values, greatly reduced the memory required to store data. These data frames were passed to the algorithm without any information about the rotation of the object corresponding to each frame. This was done to simulate the unknown-orientation aspect of the single molecule imaging process.

Three more data sets were generated by combining 2, 4, and 10 consecutive frames within the 1000 frame batches, respectively. Since the object rotated ap-

proximately 8×10^{-3} degrees between frames, we can safely assume that 10 consecutive frames are from essentially identical orientations of the object.

3.4 Pre-processing of data

Figure 3.2(a) shows the angle-averaged pattern obtained by summing over all frames. Since pixels in the outer region (e.g. top-right and top-left corner) are never obscured by the object, they provide no structural or orientational information. From the histogram in Fig. 3.2(c), a photon count of 17,800 was chosen as the cutoff to define a mask of pixels as shown in Fig 3.2(b). Only the green pixels are used to determine the orientations of each frame of data. In the remainder of the paper, these pixels are referred to as relevant and the others as irrelevant.

Another mask (red in Fig. 3.2(b)) is used to exclude the pixels in the gaps between detector tiles. If this is not done, the algorithm naturally interprets these gaps as coming from an infinitely opaque structure obstructing the view in every frame. Since there are no photons here, the exact attenuation caused by this structure is undefined except that it is above a certain level. In addition to these pixels, this mask also includes 7 "hot" pixels that were malfunctioning and erroneously record extremely high count rates. This mask of ignored pixels is shown in red in Fig. 3.2(b). The pattern is not symmetric or smooth because of statistical noise, small variations in pixel gain, detection efficiency and small errors in pixel offsets. The lack of symmetry and smoothness present no problem for data reduction and

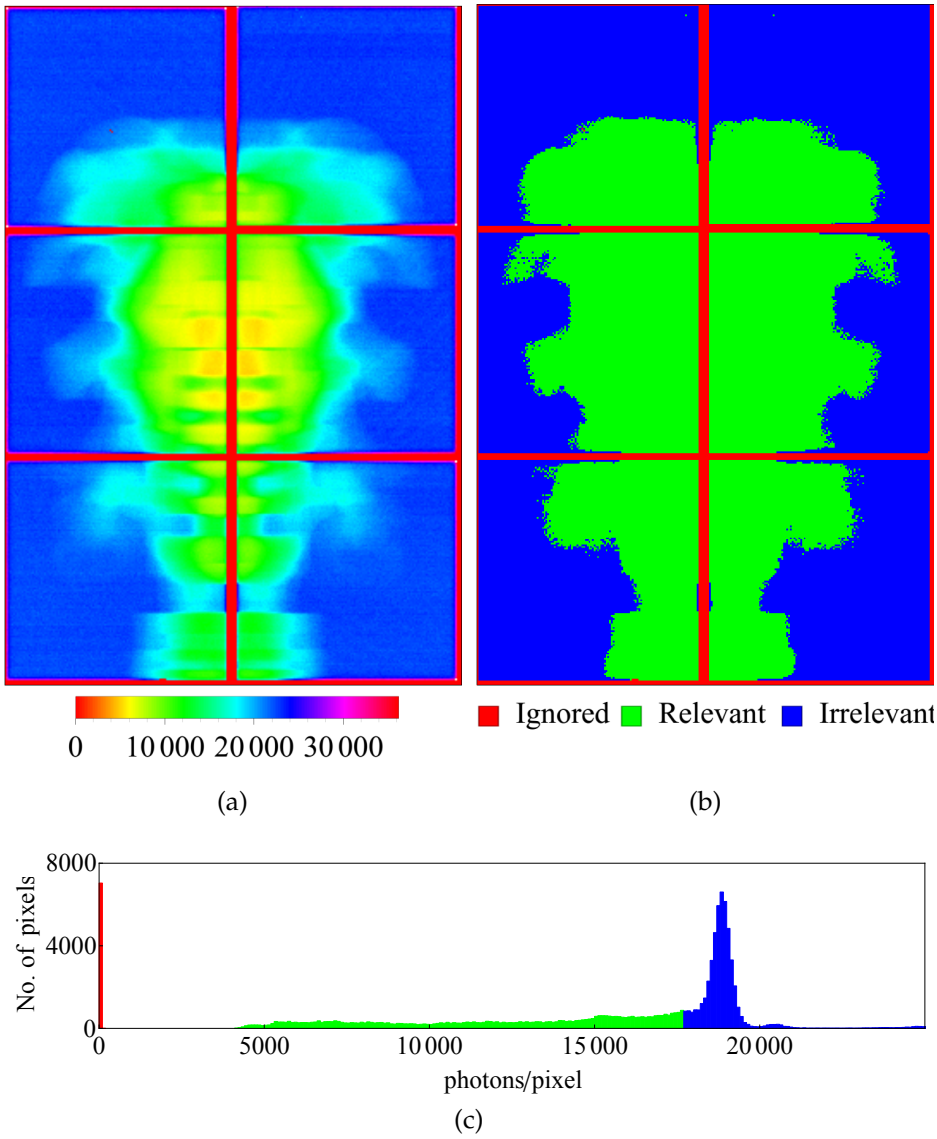


Figure 3.2. (a) Angle-averaged pattern produced by summing over all 15,650,615 frames in the 99 photons/frame data set. The numbers in the legend refer to photon counts. (b) Mask representing relevant (green), irrelevant (blue) and ignored (red) pixels (details in Section 3.4) (c) Histogram of photon counts with a cutoff value for relevant pixels at 17,800 photons.

the algorithm is generally robust in this regard.

3.5 Reconstruction algorithm

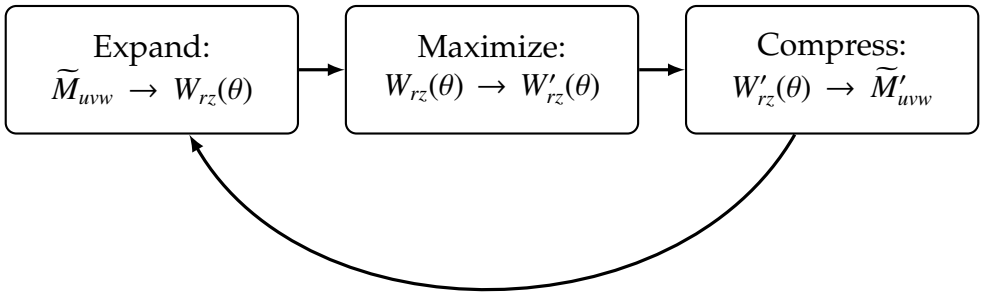


Figure 3.3. Flowchart of EMC reconstruction algorithm applied to this system including the transformations performed in each step.

The imaging process can be thought of as unknown-angle tomography at very low signal. This problem naturally splits into two parts, tomography and determination of angle. The projection-slice theorem was used to tackle the first and an iterative expectation-maximization (EM) based algorithm for the second.

Since the projection-slice operation is applied partly in Fourier space, it is convenient to have the iterate be a 3D Fourier space model $\widetilde{M}_{uvw}$. The 3D inverse Fourier transform of $\widetilde{M}_{uvw}$

$$\text{3DIFT}[\widetilde{M}_{uvw}] = M_{xyz} \tag{3.1}$$

gives the attenuation, by $\exp(-M_{xyz})$, of x-rays passing through voxel xyz of the object.

The relevant pixel mask is used to generate the initial random model. Consider the three-dimensional object generated by rotating this mask about the axis of rotation. Since we assume that all pixels outside the relevant region are never obscured for any angle, this 'rotated-mask' object acts as the support for the target object. Thus, voxels inside this object are assigned a random number uniformly in the range [0,1] and the voxels outside are zeroed. This 3D array is zero-padded perpendicular to the rotation axis to reduce interpolation errors and Fourier transformed to generate the initial random Fourier model.

We express our algorithm in the Expand-Maximize-Compress (EMC) framework of [15]. Starting with the initial random model, in each iteration, these three operations are applied to it to generate the updated model. Here the Expand and Compress steps represent transformations between the 3D attenuation model in Fourier space and the collection of 2D real-space intensity attenuation patterns, for uniformly sampled object-rotation angles. We will refer to the latter as tomograms. The Maximize step generates updated tomograms which increase the likelihood of the data being generated from the model. Thus, in each iteration, we generate the tomograms from the current model (Expand), update the tomograms (Maximize), and combine them to generate the new model (Compress).

Figure 3.4. The Expand step generates tomograms, $W_{rz}(\theta)$, for many different discrete orientations θ from the 3D Fourier space model $\widetilde{M}_{uvw}$.

3.5.1 Expand

The tomograms are generated using the projection-slice theorem augmented by exponentiation of the projections. We use linear interpolation to generate slices $\widetilde{T}_{\rho w}(\theta)$ of $\widetilde{M}_{uvw}$ passing through the axis $(u, v) = (0, 0)$ for a large number of uniformly spaced orientations, θ. To avoid interpolation errors, we oversample the Fourier space model. Thus, if there are $U \times W$ pixels in the detector and the rotation axis is along the w-axis, the iterate has $(\sigma U)\times(\sigma U)\times W$ complex voxels, where σ is the oversampling factor.

We inverse Fourier transform the slices $\widetilde{T}_{\rho w}(\theta)$ to generate the projected attenuations $T_{rz}(\theta)$ and then generate the intensity models by applying the formula,

$$T_{rz}(\theta) = \text{2DIFT}[\widetilde{T}_{\rho w}(\theta)] \tag{3.2}$$

$$W_{rz}(\theta) = f.\exp[-sT_{rz}(\theta)] \tag{3.3}$$

Here, f represents the unattenuated intensity and s represents a scale factor explained below.. Since we assume that the irrelevant pixels are unobstructed in any orientation, we can obtain f from the mean photon count in these pixels.

Before we apply the Maximize step on $W_{rz}(\theta)$, we must determine if we have the right overall scale for $T_{rz}(\theta)$. From the data, we know the mean number of

photons/frame, $\langle \sum_{rz} W_{rz}(\theta) \rangle$, where the angle bracket denotes averaging over all orientations θ. However, due to the exponentiation in Eq. (3.3), we cannot use this to directly calculate $\langle T_{rz}(\theta) \rangle$. We also observe that the algorithm is prone to a scaling instability in the first few iterations where the values in $T_{rz}(\theta)$ explode, leading to underflows in the intensity due to Eqn 3.3. This is avoided by numerically determining the correct scale every iteration. We use the secant method [27] to solve

$$0 = \langle W_{rz}(\theta) \rangle - f.\langle \exp[-sT_{rz}(\theta)] \rangle \tag{3.4}$$

for the unknown scale factor s.

3.5.2 Maximize

In this step we find the updated intensities $W'_{rz}(\theta)$ using expectation maximization. Due to the low signal count per pixel, we expect the probability of a photon incident on a pixel to be governed by Poisson statistics. At each pixel, the intensity model value gives the mean of this Poisson distribution. Thus, the likelihood of a frame of data d (K_d) coming from $W_{rz}(\theta)$ is given by,

$$\ell_d(\theta) = \prod_{rz} \frac{W_{rz}(\theta)^{K_{d,rz}} \exp[-W_{rz}(\theta)]}{K_{d,rz}!} \tag{3.5}$$

$$\Rightarrow p_d(\theta) = \frac{\ell_d(\theta)}{\sum\limits_{\theta} \ell_d(\theta)} \tag{3.6}$$

where $p_d(\theta)$ is the probability obtained by normalization and $K_{d,rz}$ is the number of photons at pixel (r, z) in frame d. The $K_{d,rz}!$ factor in the denominator of Eq. 3.5 can-

cels out in the calculation of $p_d(\theta)$. To increase the effectiveness of our probability assignment, we only consider photons in relevant pixels.

Using these probabilities we calculate the updated intensities, $W'_{rz}(\theta)$, which maximize the log-likelihood of generating the data,

$$\log[Q(W'_{rz}(\theta))] = \sum_d \sum_\theta \sum_{rz} p_d(\theta)[K_{d,rz} log(W'_{rz}) - W'_{rz}] \tag{3.7}$$

Rearranging the sums and maximizing with respect to W'_{rz}, we get

$$W'_{rz}(\theta) = \frac{\sum_d p_d(\theta) K_{d,rz}}{\sum_d p_d(\theta)} \tag{3.8}$$

For two reasons explained below, we apply an "inertia" factor α in the update rule for $W_{rz}(\theta)$. We apply an update rule,

$$W'_{rz}(\theta) \leftarrow \alpha W_{rz}(\theta) + (1 - \alpha) W'_{rz}(\theta) \tag{3.9}$$

Thus, $W'_{rz}(\theta)$ has a contribution from the previous iteration.

In the first few iterations, where the iterate changes rapidly, a high value of α prevents $T_{rz}(\theta)$ from exploding, leading to underflows in the intensity calculation. Secondly, it also provides an additional handle on the rate of convergence of the algorithm. In the current geometry, the various slices $\widetilde{T}_{\rho w}(\theta)$ do not overlap each other except at the rotation axis. This means that there is only a weak constraint for successive slices to correspond to successive rotation angles. Thus, there are near solutions which have arbitrary jumps in angles in successive slices. Indeed, this is what we observe we converge to when we have relatively high signal. However,

if we slow down the rate of convergence using a high inertia factor, we reliably converge to the right solution. We do not expect this to be an issue if we have full 3D rotation as the various slices strongly overlap. In that case, the Expand and Compress steps should together provide a strong constraint which impose the correct ordering on the slices.

3.5.3 Compress

This is the inverse of the Expand step. First, we generate the attenuation projections $T'_{rz}(\theta)$ by taking the negative logarithm of the updated tomograms,

$$T'_{rz}(\theta) = -\log(W'_{rz}(\theta)/f)) \tag{3.10}$$

The ignored pixels from the panel gaps and "hot" pixels are not updated. These updated projections are then zero-padded and Fourier transformed to get the updated slices $\widetilde{T'}_{\rho w}(\theta)$. We then use linear interpolation to generate the updated 3D model.

3.6 Results

Data was taken with 99.5 mean photons per frame. Of these, only 37 photons were incident in the relevant region of the detector as defined in Section 3.4. Using the mean incident flux calculated from the irrelevant pixels, we can determine that on

Table 3.1. Parameters of four data sets analyzed. The last three were generated by combining 2, 4, and 10 successive frames of the first data set. Relevant pixels refer to the green region in Fig. 3.2(b). Only the photons in this region have orientational information about the data frames.

	Total photons/frame	No. of frames (millions)	Relevant photons/frame	Absorbed photons/frame
1	99.5	15.6	37.0	3.9
2	198.7	7.8	73.8	7.8
3	397.2	3.9	147.6	15.6
4	991.6	1.6	368.4	39.1

average around 4 photons were absorbed by the object per frame. As mentioned in Section 3.3, frames were grouped by either 1, 2, 4, or 10 consecutive frames into 4 data sets. After this initial grouping, no information on the angular position of the combined frames was passed to the reconstruction algorithm. Table 3.1 lists the properties of all four data sets.

Figure 3.5 shows reconstructions from the four data sets along with a set of high flux, static projections. In all four cases, the only thing that changes is the signal per frame. The total number of photons and the object is unchanged. We can clearly see that the quality of the reconstruction improves as we increase the mean number of photons/frame. With 99 photons/frame, the algorithm is unable to determine even the gross shape of the object. The finer, low-contrast features are reconstructed with more and more accuracy as we increase the number of frames combined. In the bottom row, the circle shows the location of the extra upper arm, which was attached only on one side. The oval shows the asymmetry in the two lower arms due to one of them holding a plastic dumbbell.

Note that even with 10 frames combined, there are still only 7.9×10^{-3} photons/frame/pixel in the relevant region of the detector.

3.7 Methods

For all four data sets, an oversampling factor, σ, of 2 was chosen. Thus, the Fourier-space iterate had dimensions $396 \times 532 \times 532$.The angular range was divided into 200 discrete orientations and the inertia factor was chosen to be 0.8 in each case. A cutoff count of 17,800 was used to generate the relevant pixels mask. There were 46,890 relevant pixels and 8,670 ignored pixels.

Due to the large size of the problem, a parallel algorithm was required. The most time-consuming step in each iteration was implementing Equation 3.5. The parallelization was applied over the θ index i.e. each process was assigned a part of the angular range for which to calculate $p_d(\theta)$. Reconstructions were performed at the LCLS cluster at SLAC. With 120 processes on 10 nodes, an iteration took 2-3 minutes and the algorithm took 30–50 iterations to converge. To test for convergence, reconstructions were performed multiple times with different random starts and they yielded the same result.

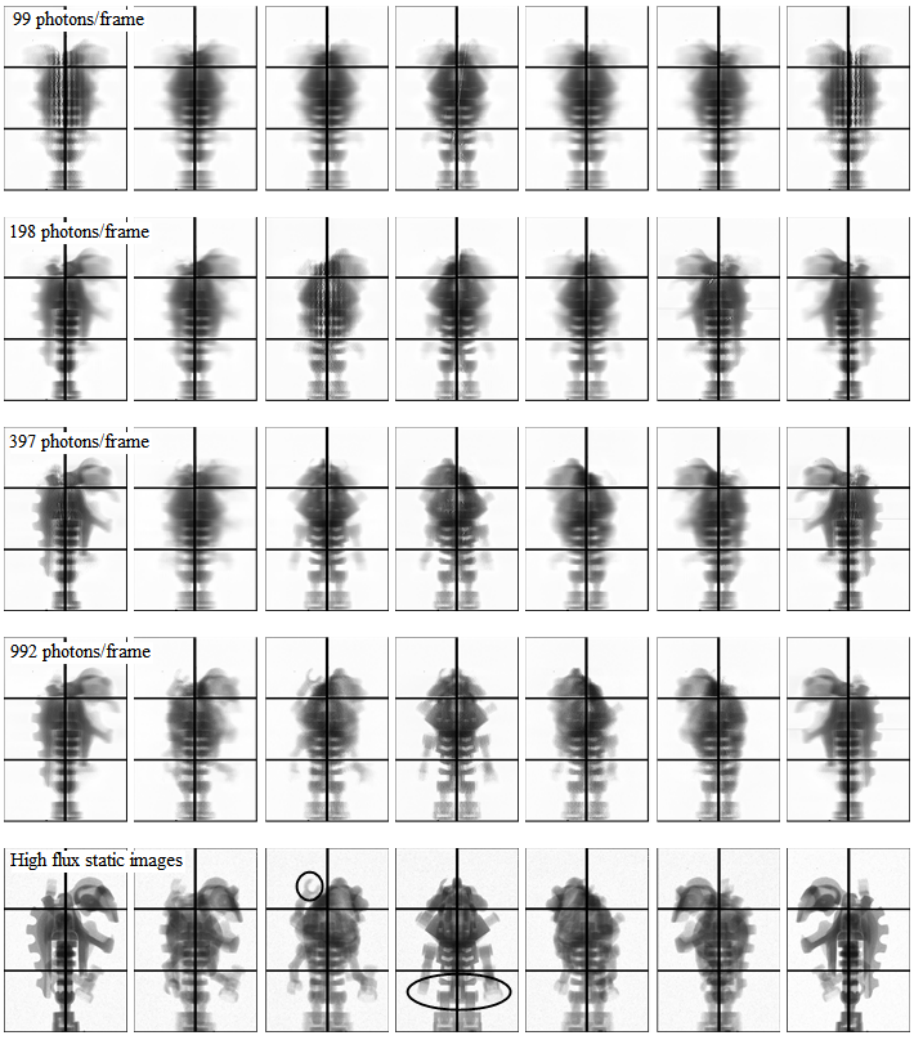


Figure 3.5. The first four rows show the projected x-ray transmission intensities through the reconstructed object for $\pi/6$ rotation intervals in $[0,\ \pi]$. Reconstructions of the object were obtained with data sets of 99, 198, 397, and 992 photons per frame, respectively. The total number of photons in each data set was 1.56×10^9. Details about the data sets are given in Table 3.1. The bottom row shows for comparison static radiographs of the object which were acquired at the same angles at high signal levels. Some fine, low-contrast features are circled. All images are scaled such that white and black colors represent no and complete attenuation respectively.

3.8 Conclusions

We view these results as an extension of the 2D reconstructions performed previously with similarly sparse data in [19]. The in-plane rotation axis in the present study meant that we had to reconstruct a 3D object. Any given frame in this experiment did not have the full structural information. Due to the large size of the object, a non-linear attenuation model had to be used. Also, we needed a parallel code running on a cluster to perform the reconstruction in a reasonable time. With this experiment, we are one step closer to demonstrating the reconstruction of the 3D intensity of a biomolecule or nanocrystal in conditions of very low signal.

There are further avenues that could be pursued to improve the quality of the reconstruction: 1) One could take into account the small, but finite divergence of the beam. This would make the analysis in the expand and compress steps similar to fan-beam tomography. 2) The data at the edge pixels could be modified to reflect their larger size. 3) The axis of rotation was assumed to be aligned along the middle of the detector. This could estimated more accurately. 4) Our criterion for the relevant pixel mask was chosen for simplicity (a hard cutoff on the number of photons/pixel). This could be further refined to maximize the capacity to determine the orientation of a frame.

In the process of constructing a 3D proof-of-principle experiment for biomolecule imaging with x-rays, our experiment closely approaches the setup of cryo-electron microscopy (cryo-EM) experiments for biomolecules [21]. In that

case, we also have tomography with unknown orientations at very low signal-to-noise. However, there are a few differences. The cryo-EM reconstructions do not have the assistance of a rotation axis, and so have molecules in all orientations in SO(3). In addition, they must also fix translational alignment in their individual frames. Finally, the x-ray data here is in the low signal regime of Poissonian statistics, while the noise model for cryo-EM is not as simple.

The results in Fig. 3.5 suggest that there is a minimum number of photons/frame needed to determine the structure of the object. A similar feasibility criterion was found in the EMC simulations for single molecule diffraction imaging [7, 15]. In this case, the criterion depends on the particular object. More specifically, the more attenuation there is far from the rotation axis, the easier it is to assign orientations to frames. Also, a higher contrast object with the same shape would be easier to reconstruct.

The goal of reconstructing a single biomolecule or microcrystal needs a few more steps. First, we need to demonstrate a reconstruction with randomly-oriented diffraction data. Secondly, there would be sources of background we have not included, such as air-scatter, which would make the reconstruction more challenging.

Acknowledgments

Research on the development and application of x-ray detectors is supported by DOE Grant DE-FG02-10ER46693, the Keck Foundation, and CHESS. CHESS is supported by NSF and NIH-NIGMS under NSF Grant DMR-0936384. The data analysis work is supported by DOE Grant DE-FG02-11ER16210. The sample used to demonstrate tomographic reconstruction was graciously provided by Russell K. Philipp.

CHAPTER 4

CRYSTALLOGRAPHY WITH SPARSE DATA

This work was carried out with Hugh Philipp, Mark Tate, Jennifer Wierman, Veit Elser and Sol Gruner. It is in preparation.

4.1 Abstract

X-ray serial microcrystallography involves collection and merging of frames of diffraction data from randomly oriented protein microcrystals. The number of diffracted x-rays in each frame is limited by radiation damage, and this number decreases with crystal size. The data frame is said to be sparse if too few x-rays are collected to determine the orientation of the microcrystal. It is commonly assumed that sparse crystal diffraction frames cannot be merged, thereby setting a lower limit to the size of microcrystals that may be merged with a given source fluence. The EMC algorithm [15] has previously been applied to reconstruct structure of sparse noncrystalline data of objects with unknown orientations [19, 1]. Here it is shown that this method may be extended to serial microcrystallography. As a proof-of-principle, we demonstrate reconstruction of the 3-dimensional diffraction using sparse data frames from a small molecule (1.35 kDa) crystal. The results indicate that serial microcrystallography is in principle not limited by the fluence of the x-ray source and collection of complete data sets should be feasible at, e.g., storage ring x-ray sources.

4.2 Introduction

Serial microcrystallography was developed as a way to take advantage of the high fluence provided by x-ray free electron lasers to image small microcrystals (c.a. $1\mu m^3$ or smaller) [5, 4]. Due to the short time-duration of the pulse (< 50 fs), the principle of "diffraction-before-destruction" is applicable where the pulse outruns most of the radiation damage. This allows the capture of relatively damage-free snapshot diffraction patterns. A large number of these patterns are captured by flowing a stream of crystals past the beam. Enough photons are scattered in this interval to allow indexing algorithms [28] to determine the orientation of individual frames and to generate the 3-dimensionsl (3D) intensity distribution of the diffraction.

This approach was reproduced at a synchrotron source [23] with larger crystals ($135\mu m^3$) and the same indexing method. However, with micron-sized crystals, around 100 times fewer photons would be scattered. In this case, there would be too few photons in a single frame to allow indexing of Bragg spots. The data would have the sparse nature of [19] and [1], where it is impossible to recover the orientation of a single frame by itself. Fortunately, as in those cases, we show that one can apply the EMC algorithm [15] to simultaneously assign orientations and solve for the 3D intensities.

To simulate the sparse frame conditions from a $1\mu m^3$ crystal at a storage ring synchrotron source, we have performed an experiment with a large 1-mm sized

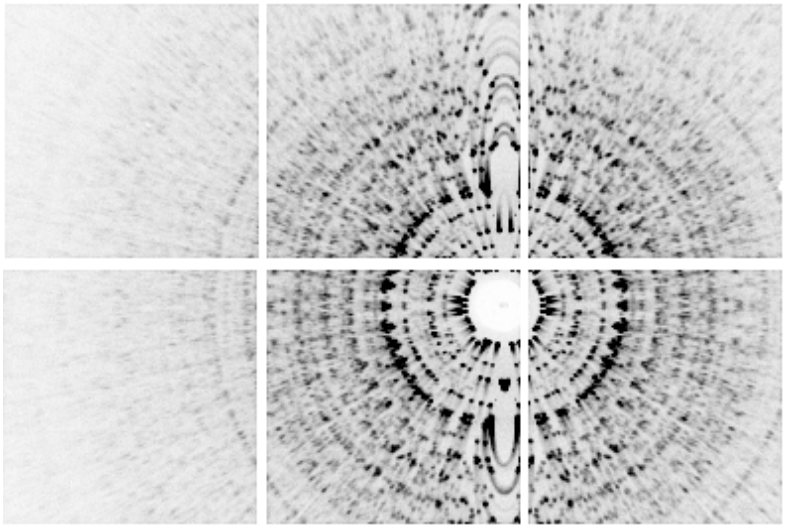


Figure 4.1. Angle-averaged pattern produced by summing over all frames in the low-fluence data set. The direct beam goes through the center of the beamstop and the rotation axis is vertical. Note the radial streaks caused by non-monochromaticity of the beam due to bremmstrahlung. These streaks form arcs near the vertical rotation axis due to the curvature of the Ewald sphere.

crystal with standard laboratory x-ray source with diffraction images recorded at a high frame rate. Each frame was acquired with the crystal in an arbitrary orientation around a single rotation axis. Even with a signal level as low as 4.8×10^{-3} photons/pixel/frame, we demonstrate successful recovery of orientation about the axis of rotation and reconstruction of 3D intensities. We compare our reconstruction with a high-fluence data set where the orientations were recorded. We also examine the effect of background scatter on the quality of the reconstruction and the ability to recover orientations.

4.3 Reconstruction Algorithm

A slightly modified version of the EMC algorithm [15] was used to iteratively assign orientations and reconstruct the 3D intensity distribution. One feature of this technique is that all regions of reciprocal space are treated equally. No particular preference is given to reciprocal lattice points. This is in contrast to the approach taken by indexing algorithms which emphasize the Bragg spots to the extent of ignoring the diffuse scattering. In the case of sparse data, most Bragg spots will produce no photons and some of the photons could be from non-Bragg background, making this approach impractical. A short review of the algorithm is given below.

The EMC algorithm has three steps in each iteration (Expand, Maximize and Compress). First the space of available orientations is discretized to a finite number of angles. The Expand and Compress steps convert to and from the 3D intensity distribution to the expected intensity at the detector in each of these orientations, which we call 'views'. This is done using linear interpolation along the Ewald sphere. The Maximize step uses the data frames to update the views using expectation-maximization as described below.

Once the views have been obtained for every discrete orientation, the probability of a frame being generated by a view is calculated by assuming Poisson statistics for the number of photons recorded given an intensity. Thus, if the intensity at pixel t in view r is W_{rt}, the probability of frame d with K_{dt} photons at

pixel t, being generated by view r is given by

$$P_{dr} = \frac{\prod_t e^{-W_{rt}} W_{rt}^{K_{dt}}}{\sum_r \left(\prod_t e^{-W_{rt}} W_{rt}^{K_{dt}} \right)}$$

where the $K_{dt}!$ term has been omitted as it cancels out during normalization. Using these probabilities, the likelihood maximizing updated view W'_{rt} is given by

$$W'_{rt} = \frac{\sum_d P_{dr} K_{dt}}{\sum_d P_{dr}}$$

This intuitive update rule ends up being just the weighted mean over the data frames with the weights being the probabilities calculated using the current model. These updated views maximize the likelihood of generating the data given the probabilities P_{dr} calculated from the current model. The Expand-Compress cycle is necessary to impose consistency among different views by requiring that they come from a common 3D model.

4.3.1 3D *hkl*-space

One modification to the traditional algorithm is in the choice of space for the 3D intensity distribution. The standard choice is Fourier space, where the slice representing the detector plane is the surface of a sphere passing through the origin (the Ewald sphere). Here, for reasons explained below, the best choice is *hkl*-space

where the three axes represent the fractional coordinates with respect to the reciprocal unit cell. Thus, the reciprocal lattice points lie on a cubic grid with integer spacing regardless of unit cell parameters. Unless the crystal has cubic symmetry, the detector pixels will no longer lie on the surface of sphere, but along some other surface. The pixel coordinates in this space can be calculated by using the basis vectors to determine the scaling and rotational transformation to the Ewald sphere surface.

By choosing integer *hkl*-point spacing, the center of each Bragg peak lies on a gridpoint. Thus, while interpolating in the Expand and Compress steps, equivalent Bragg peaks see the same environment. This is important because the Maximize step is sensitive to slight changes caused by interpolation errors. Another convenient by-product of this is that peak integration is standard across all crystals and meaningful comparisons of peak shapes can easily be made.

4.3.2 Initial guess

As with any iterative algorithm, the initial guess model must be specified. In this case, we assume that the unit cell basis vectors are known a priori. This can be calculated using the angle-averaged pattern obtained by summing all data frames. Using these vectors, a reciprocal lattice is constructed and a 3D Gaussian of random height is placed at each lattice point. This cubic grid of random intensities is used as the input for the first iteration.

4.3.3 Rotation group subset

In general, the set of views, r, are generated by sampling the 3D rotation group uniformly. This is done with the help of unit quaternions. However, in cases where the crystal is rotated about a single axis and the relative orientation of the axis with respect to the crystal basis vectors is known, one can sample angles about just that axis. This was done in this experiment, where the axis was determined from the high fluence dataset using the HKL-2000 software [17].

4.4 Data Collection

The sample studied was a $\sim$ 1-mm sized small molecule single crystal with chemical formula $C_{78}H_{120}Mo_2N_6O$ (mol. wt. 1.35 kDa). It was mounted on the end of a glass fiber attached to a goniometer head which allowed the crystal to be centered on the rotation axis. A rotation stage (Newport URS100) was set to rotate continuously at 0.1 degrees/s during data collection. Although the angle of rotation was known for each frame, it was not recorded or passed to the reconstruction algorithm. The crystal was illuminated by a Molybdenum K_α beam generated by a Rigaku RU-3HR rotating anode set to 30 kV 40 mA. Filtering was done using 200 μm of Zirconium foil to increase the fraction of K_α radiation. X-rays were focussed to a 0.5 mm x 0.5 mm spot using Nickel coated Franks mirrors 1 m from the sample with a beam convergence of 1 mrad. The data was recorded using the MMPAD detector [24]. Two data sets were collected, low fluence at 10 ms/frame

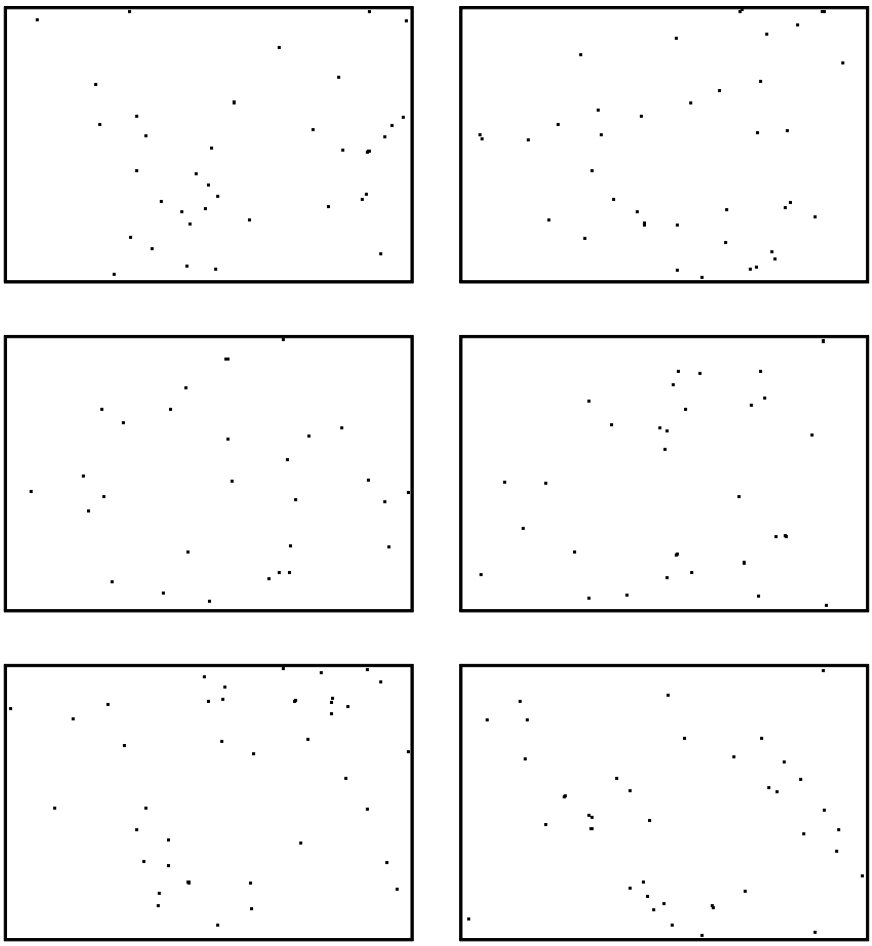


Figure 4.2. Six typical data frames obtained by collapsing 15 successive low-fluence frames together. Each frame has on average 48 photons. The locations of the photons have been emphasized to improve visibility.

and high fluence at 500 ms/frame. The low fluence data set was taken in groups of 1000 consecutive frames with a time delay between sets to allow the frames to be written to disk.

The data was then thresholded and photon counts were obtained using a procedure similar to that employed in [1]. In the low fluence data set there were 4.3 million frames with an average of 3.2 photons/frame. Since the crystal rotated only 0.001 degrees between two successive frames, multiple data sets were prepared by combining successive frames within a batch. Table 4.1 lists the details of the different data sets. Figure 4.2 shows the first six frames from the 100 frame data set.

Figure 4.1 shows the angle-averaged pattern obtained by summing over all data frames. The radial streaks near Bragg spots are caused by the polychromaticity of the beam. The arcs near the rotation axis are caused by the these streaks intersecting the curved Ewald sphere. Since the exact shape of the arc is very sensitive to the rotation axis, a region of the image within 11 pixels of the rotation axis was not used in the calculation for P_{dr}.

4.5 Results

The high fluence data set with known orientations was used to generate a reference 3D intensity model. This was compared with the reconstructions from differ-

Table 4.1. Table showing properties of various data sets generated by collapsing successive frames. Before collapsing, the frames were in batches of 1000 contigous frames with gaps. There were also some rejected frames due to errors. The number of iterations required for convergence depends upon the random start so the numbers given here are approximate and are used to highlight the trend.

Collapsed frames	# of frames	photons/frame	Iterations to converge
1	4,321,197	3.22	–
10	434,420	32.00	–
15	290,541	47.85	2200
100	44,221	314.41	400
200	22,321	622.88	250

ent low fluence data sets by comparing the Patterson maps, which were generated as follows. First, the intensities were integrated in a small sphere about every *hkl* point. The 3D *hkl* grid of intensities was then inverse Fourier transformed to generate the electronic density auto-correlation function, which is the Patterson map. Fig. 4.3 shows the comparison of the maps for one particular data set (314 photons/frame). We consistently observed that if the algorithm converged, it produced very similar maps. Thus, convergence was taken to be an indicator of a successful reconstruction. Here iterative convergence occurs when the iterate stops changing form one iteration to the next. Some convergence plots are shown in Fig. 4.4.

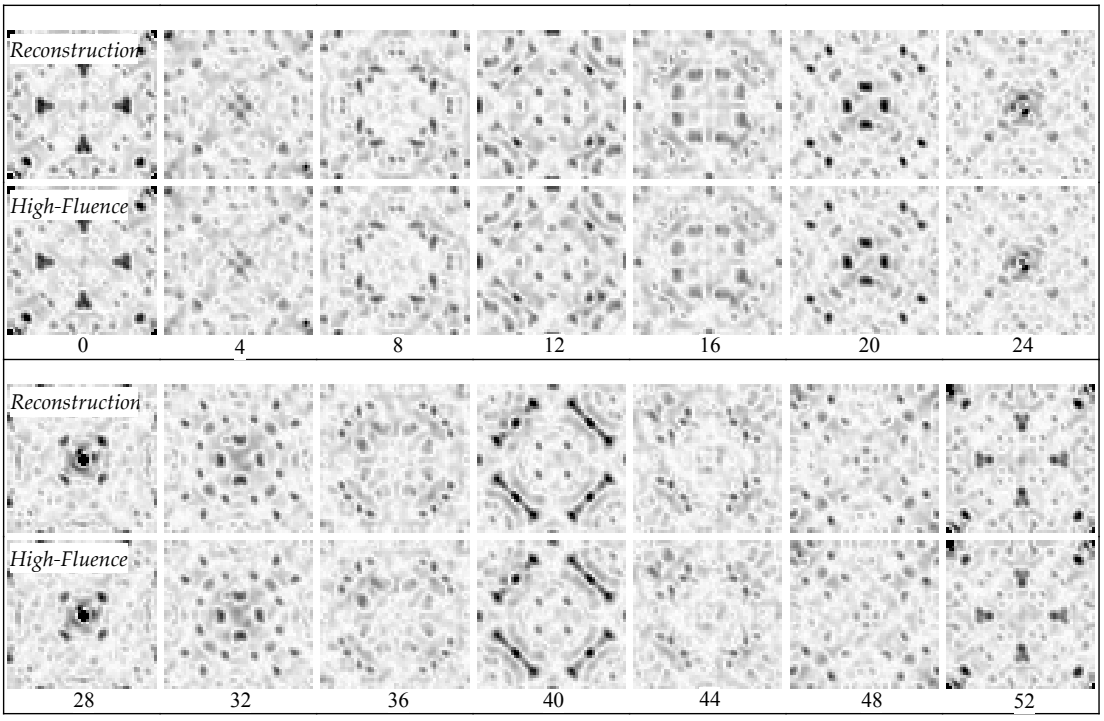


Figure 4.3. Comparison of slices through the 3D Patterson maps generated from the high-fluence data set and a low-fluence (48 photons/frame) reconstruction. The map was $53 \times 53 \times 53$ voxels in size and every 4th slice is shown here with the slice number shown below each pair.

4.5.1 Dependence on photons/frame

As mentioned in Section 4.4, the crystal rotated 10^{-3} degrees over one frame. This allowed us to collapse successive frames as they came from almost identical orientations. Using this method, we could study the effect on reconstruction quality of number of photons/frame while keeping other parameters the same. One effect of decreasing the number of photons/frame was that it took more iterations to reach convergence. For less than 48 photons/frame, the reconstruction did not converge

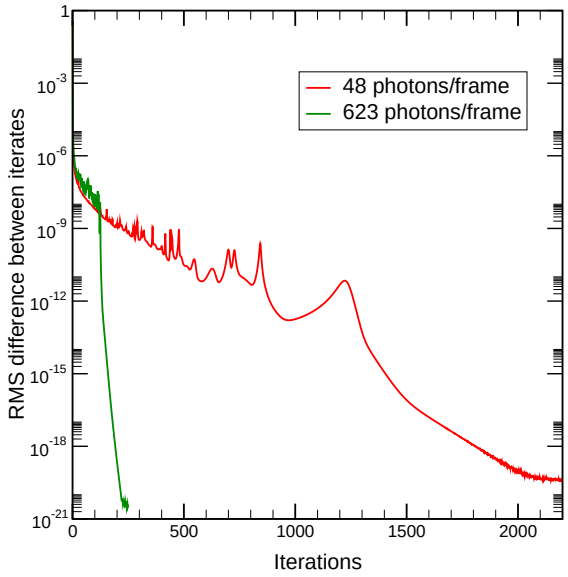


Figure 4.4. Plot showing the difference between successive iterates as a functino of iteration number. The scale on the y-axis is arbitrary, but the lower limit is near machine precision. Two data sets are shown with 15 collapsed frames (48 photons/fame) and 200 collapsed frames (623 photons/frame). The sparser data set takes much longer to converge and the slope of the curve in the last few iterations is strongly related to the number of signal photons/frame.

at all. Above this threshold value, the reconstruction quality was independent of number of photons/frame. However, it took many more iterations, as can be seen in Fig. 4.4 and Table 4.1. This is consistent with the observations in simulations with speckle intensity patterns [15]. The threshold value itself is lower because of the different distribution of the intensity in this case (concentrated in Bragg peaks as opposed to large, smooth speckles).

4.5.2 Addition of background

Due to the large size of the crystal and the fact that, being a small molecule, it was not hydrated, there was relatively little background scattering compared to the Bragg spots. To study the effect of uniform background on the quality of the reconstruction, additional photon counts were added with a Poisson distribution of uniform mean to each data frame. Except in the cases of extreme background, there is no effect on the orientation recovery. The weak, highest-resolution peaks are lost as they are drowned out by the noise in the background. This is an unavoidable aspect of crystallography.

To demonstrate this, the ratio of average intensity per voxel in the neighborhood of a Bragg point to the average intensity in the diffuse region is plotted versus reciprocal length, q, in Fig. 4.5. If this ratio is close to 1, the Bragg peaks do not stand out over the background. As the plot shows, even with high background, the strong, low-resolution peaks are successfully recovered. However, as expected, the weak, high-resolution peaks are lost. The sharp troughs in the ratio at lower resolution are caused by the radial streaks in the intensity due to the polychromatic beam. These streaks are not integrated and thus increase the average diffuse intensity in the q-shells just outside a shell with strong Bragg peaks.

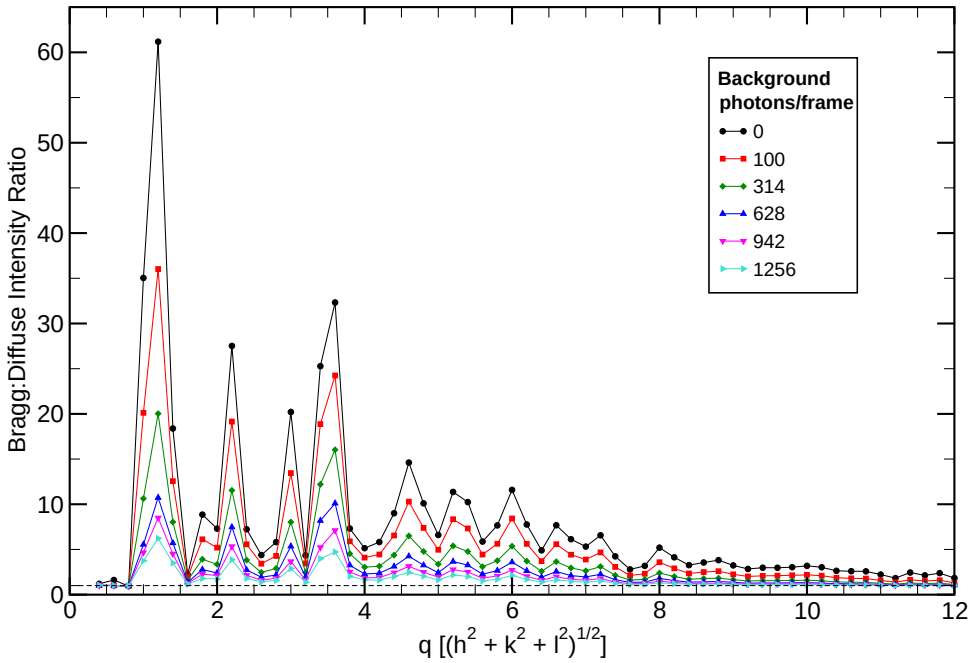


Figure 4.5. Plot of Bragg to Diffuse intensity ratio vs q for various amounts of additional background photons/frame. A high ratio indicates that the orientations have been correctly identified and most of the intensity is in Bragg peaks. There were 314 photons/frame in the base data set. Even with 400% background, the low resolution peaks could be resolved as seen the 1256 photons/frame plot.

4.6 Conclusions

We have shown that the 3D diffraction intensity distribution can be calculated from a large number of sparse data frames, each with unknown orientation. This results bodes well for the possibility of performing serial crystallography with micron-sized or smaller microcrystals at synchrotron sources. Successful reconstruction was shown for a signal level as low as 47 photons/frame.

We also observe that the addition of relatively high levels of uniform back-

ground (400%) does not affect orientation recovery. This is important as some base level of background scattering is unavoidable with protein crystals due to the solvent internal to the crystal. However, it does reduce the resolution as higher order peaks get drowned out by the background. Thus, it is desirable to lower it to get the best resolution, as is the case in conventional crystallography. This would require minimizing the amount of material in the beam path. Fortunately, it is possible to reduce background to insignificant levels by appropriate x-ray optics, vacuum paths and graphene windows surrounding the crystal stream [29]. For example, one can envision flowing a filtered set of uniformly sized microcrystals down a minimally sized tube equipped with graphene x-ray windows in an otherwise totally vacuum environment. In cases where the source fluence is low and exposure times are long, some degree of Brownian rotation is desirable as it reduces partial Bragg reflections. If the exposure time is longer, fast framing detectors [13, 12] can be used to artificially restrict the net degree of angular diffusion over an exposure average rotation angles.

One feature of the serial crystallography experiment not replicated here is the collection of data from all orientations in three dimensions. Reconstruction from the full rotation group was studied in simulations in [15] for aperiodic structures with speckle intensity distributions. There, it was shown that the number of photons/frame required for successful reconstruction grows only logarithmically with number of orientational samples. Although, the total number of photons required for a complete data set with good signal-to-noise ratio and good resolution will be higher than what was collected here, the fluence available at 3rd genera-

tion x-ray sources is many orders of magnitude higher than was the case in this experiment. This suggests that sub-micron, room-temperature serial microcrystallography should be feasible. Experiments to examine this prediction will, no doubt, be performed.

Acknowledgements

Research on the development and application of x-ray detectors is supported by DOE Grant DE-FG02-10ER46693, the Keck Foundation, and CHESS. CHESS is supported by NSF and NIH-NIGMS under NSF Grant DMR-1332208. The data analysis work is supported by DOE Grant DE-FG02-11ER16210. We thank Emil Lobkovsky for kindly providing the sample.

APPENDIX A

DETAILS FOR CRYSTALLOGRAPHIC DATA RECONSTRUCTION

Here is the analysis methodology for the crystallography experiment. The details about the experiment and the final results are discussed in Chapter 4. A weak x-ray beam was directed at a crystal and a large number of sparse data frames were collected without recording the orientation. The aim was to determine the orientations and reconstruct the 3D crystalline intensity distribution. This was done using the EMC algorithm [15] which has been described in Chapter 1. A detailed discussion of the implementation of the algorithm in this particular case follows.

A.1 Pre-processing

Before the EMC algorithm can be implemented, some things need to be designed. The first is the organization of the data. The principal considerations are to minimize disk usage and maximize ease of access. Secondly, the 3D model of the iterate must be chosen. Associated with this is the mapping of pixels to this space. And finally, the rotation group must be sampled, either uniformly in 3D, or about a known rotation axis. The last two together define the Expand and Compress steps.

A.1.1 Data storage and access

Since the data was so sparse, there was a need for a new format to store the data in a form analogous to a sparse array. A sparse array is stored in memory by noting the address and contents of all non-zero entries. As most entries are zero, this saves on a lot of memory. In this case, there are even more savings to be made because the vast majority of non-zero counts are ones. Thus, the ideal choice is to store the locations of all pixels with one photon and the location and count of those rare pixels with more than one. To aid in parsing and processing the data, for each frame there were two numbers, denoting the number of single photon pixels (`num_ones`) and the number of multiple photon pixels (`num_multi`).

While the total size of the data set was not large in this case (50 MB), this would not be true in general. If the data file would become too large, it could not be held in memory and would require parsing in batches from disk. Thus, a system that allowed one to flexibly pick frames N to $N + M$ for any values of N and M was needed. This was achieved by placing the `num_ones` and `num_multi` counts at the start of the file for all frames. After this, the pixel numbers for the singles (`place_ones`), the multiples (`place_multi`) and finally the multiple-pixel counts (`count_multi`) were written sequentially without any delimiters for frame boundaries (Fig. A.1). While parsing the data, the frame boundaries can be easily determined from the `num_ones` and `num_multi` counts at the start of the file. It is also possible to calculate exactly what memory locations in the file a given frame number would be found with minimum effort. One disadvantage

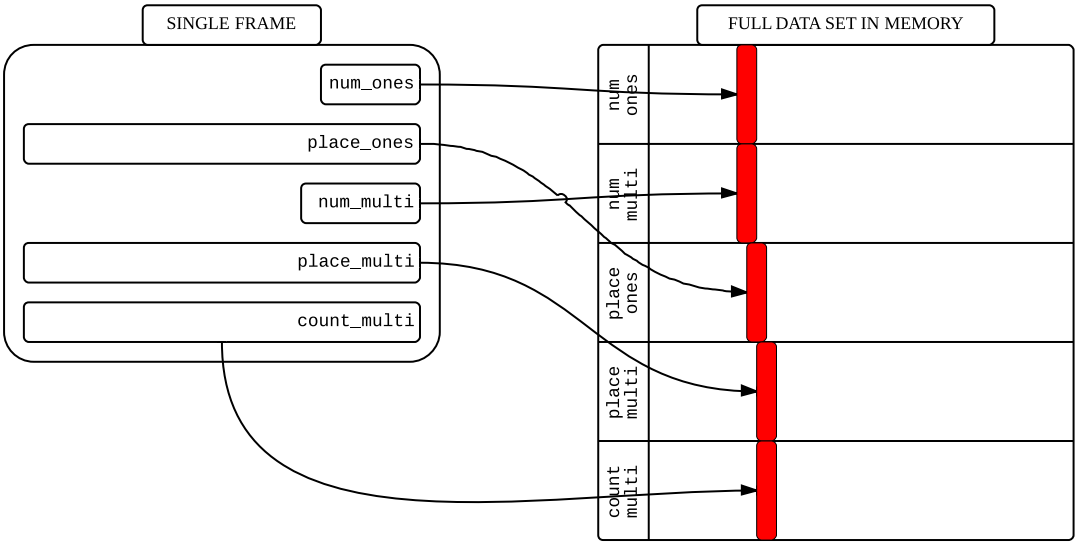


Figure A.1. Schematic description of data storage on disk. The information about a frame is not stored contiguously. Each element of the data is grouped together without frame delimiters. The first two items (`num_ones` and `num_multi`) can be used to figure out the address in memory of the other elements.

of this format is that it is inflexible to the addition or removal of frames, as compared to a system where all the information about a frame is stored contiguously. However, except for some specific cases like in-situ reconstruction, this was not foreseen to be a major driving force.

After the data was collapsed and converted to the above-mentioned format, some preprocessing was required before it could be fed to the reconstruction code. One simple step was to generate a histogram of photons/frame. Due to experi-

mental vagaries, there are always a few frames which have too high a count. These should be rejected and a blacklist file was made which told the reconstruction to skip over those frame numbers. Another easy, but very useful measure, was to examine the angle-averaged pattern, i.e. the view of the detector on adding up all the frames (colloquially called the powder pattern). This has always been very instructive in determining what are the best choices for various reconstruction parameters. In this case (Fig. A.2), the powder pattern was used to identify the arcs near the rotation axis. These are generated by radial streaks intersecting a curved Ewald sphere surface. Since the accurate positioning of these streaks is **highly** sensitive to the choice of the rotation axis, it was beneficial to not consider the pixels near the axis during orientation assignment. Another region to be excluded was the beam stop. Since the pixels in this region are blocked from receiving any intensity from the beam, any photons are stray and non-structural. These exclusions were implemented using a mask, which was accessed by the reconstruction while using the data. Since these choices about the masks are made interactively (and sometimes by trial-and-error), it was believed that a mask was the way to go rather than eliminating those photons from the data file entirely.

A.1.2 3D Model

The next choice to be made was that of the model. The algorithm updates a 3D grid of real numbers representing the intensity distribution in reciprocal space. The detector views are calculated using interpolation, which is discussed in more

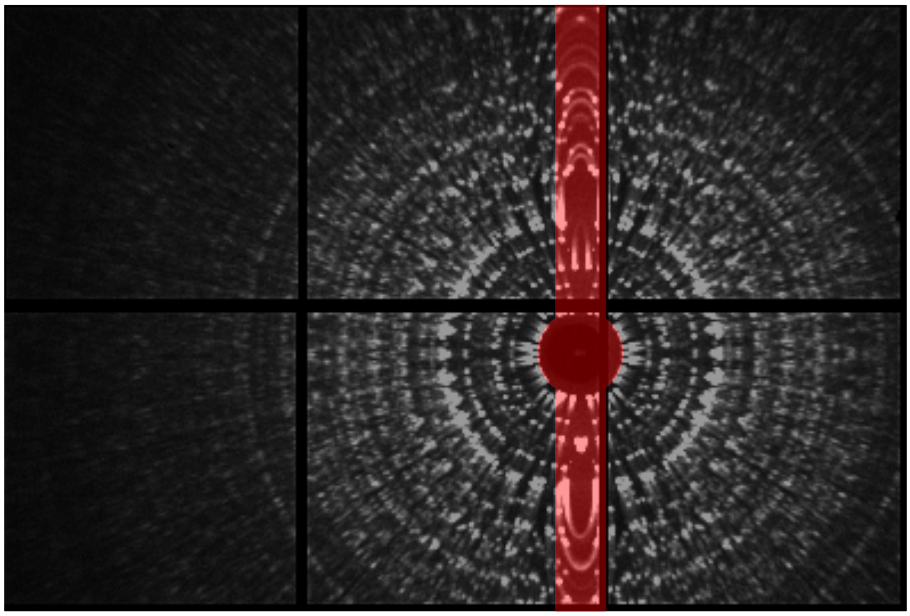


Figure A.2. Angle-averaged pattern from the crystal data set showing masked region in red. These pixels were not used in orientational assignment.

detail below. Now all interpolation produces some errors. These errors are minimized if the intensity distribution is smooth but in the case of crystalline data, the intensity is concentrated in Bragg spots which are highly non-smooth. This poses problems, specifically in the case where supposedly identical symmetry-related Bragg peaks see different interpolation environments (say one is at a grid point and the other is halfway between two points). These minor errors could become magnified by positive feedback from the reconstruction algorithm. The solution to this would be choose a basis such that all Bragg peaks lie on grid points. This was called the fractional coordinate space or the *hkl* space, referring to the fact that the axes are now not (q_x, q_y, q_z), but (h, k, l). This would require prior knowledge of the unit cell parametrs of the lattice. Fortunately, the algorithm needs these

parameters anyway to generate the initial random model. This is what was done here with a cubic lattice spacing of 5 voxels.

A.1.3 Mapping detector pixels to the 3D model

Since the EMC procedure involves comparison of data frames with views which represent the intensity on the detector, one needs to determine the pixel coordinates of each pixel in this fractional coordinate space. The pixel at the center of the beamstop is at the origin. Relative to that origin, the Fourier space coordinates of pixel (x, y) are given by

$$\frac{(x, y, D) \times D}{\sqrt{x^2 + y^2 + D^2}} - (0, 0, D) \tag{A.1}$$

where D represents the distance from the sample to the detector in pixel units. Fig. A.3 illustrates the calculation of this formula. Before the linear transformation to go from the Fourier basis to the hkl basis can be applied, the reciprocal lattice basis vectors must be calculated in voxel units. This is done using the fact that the radius of curvature of the Ewald sphere is both $1/\lambda$ m^{-1} and D voxels. Using the target lattice spacing (5 voxels in this case) and the Fourier space lattice spacing, one can calculate a scaling transformation in all three directions. If the basis vectors are not orthogonal in Fourier space, the linear transformation would have non-zero off-diagonal elements. Once these are completed, one has the 3D coordinates of each pixel in the desired space. One is now ready to calculate the views required for comparison with the data frames.

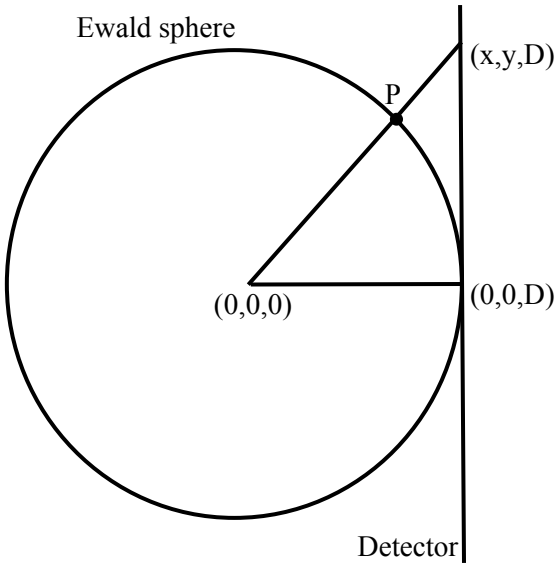


Figure A.3. Diagram illustrating the calculation of pixel coordinates in Fourier space. The detector pixel at (x, y) is mapped to the point P on the surface of the Ewald sphere. The sphere is then shifted such that the point $(0, 0, D)$ is at the origin.

A.1.4 Rotation group sampling

Three-dimensional rotation group sampling is most conveniently done using unit quaternions. This has been explained in detail in [15]. However, if the rotation axis is known, one may also just use uniformly spaced angles about the known axis. One feature of quaternion sampling is that it is never exactly uniform. This means that different views are covering different volumes of rotation space. Thus, views sampling a larger volume are scaled up using a weight factor proportional to their volumes. These factors are multiplied to the views W_{rt} before comparison with the data frames.

A.2 Reconstruction

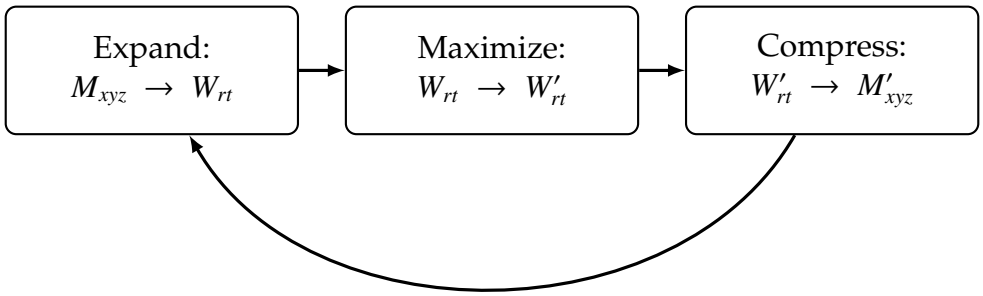


Figure A.4. Flowchart of EMC reconstruction algorithm applied to this system including the transformations performed in each step.

Figure A.4 gives a broad outline of the steps performed in each iteration. The detailed implementation of the iteration is discussed below, alongwith some methods to monitor the iterations.

A.2.1 Expand & Compress

To generate the views in the Expand step, a rotation matrix is applied to each pixel coordinate to get the location of the slice for a given orientation. In general these rotated pixel coordinates will be non-integral. The value of the pixel is calculated by linear interpolation using the 8 nearest integral neighbors forming the corners of the unit cube around the point. The weights used for the different voxels are calculated using the lever rule. As a 1D example, consider the calculation of a value at point x, $a(x)$ from the two neighbors $\lfloor x \rfloor$ and $\lfloor x \rfloor + 1$, where $\lfloor x \rfloor$ represents

the floor function.

$$a(x) = (x - \lfloor x \rfloor)\, a(\lfloor x \rfloor + 1) + (\lfloor x \rfloor + 1 - x)\, a(\lfloor x \rfloor) \tag{A.2}$$

This can be done quickly and independently for each orientation. Currently OpenMP is used for this calculation. A GPU can also be used, however the time taken to transfer the view arrays from GPU to CPU memory is the slowest step. One possible solution is to perform all operations in the GPU, including the maximization.

The inverse operation of view generation is view merging. Here, the updated views are merged together to obtain the updated 3D model. The first step in this case is again the determination of rotated pixel coordinates and the 8 nearest integral neighbors. The weights are obtained using the lever rule as before and the weighted value is added to each of the 8 neighboring voxels. Another 3D array is used to hold the weights which are added to those voxels. After all views have been merged, the model is divided by the interpolation weights. This way at each model voxel, the value represents the weighted mean of all rotated pixels which were in the neighborhood. These two steps together define the Expand and Compress steps of the EMC cycle.

A.2.2 Implementation of Maximize

The most time consuming step in the iteration is the comparison of views with data frames to calculate the probability distribution P_{dr}. The signal level in a view

is determined by the average photon count per frame. Since the intensities represent the mean of the Poisson distribution from which the counts are sampled, they must themselves be scaled such that the total intensity in a frame is this average count.

Thus, first each view is generated and the scaling factor is calculated. Since view generation is fast and trivially parallelizable, the views are not stored in memory. They are regenerated when used to calculate P_{dr}. Here is the pseudocode to calculate the factor `rescale`.

```
for r < R do
    generate_view(r)
    sub_total[r] += calculate_total(r)
    total += u[r] / R
end for
rescale = photons_per_frame / total
```

Here `u[r]` $= \sum_t W_{rt}$ will be used below.

Now that the views have been calculated and rescaled, they can be compared with the data and updated. As a reminder, the probability of frame d having orientation r is,

$$P_{dr} = \frac{\ell_{dr}}{\sum_r \ell_{dr}} \tag{A.3}$$

where the likelihood, ℓ_{dr} is given by

$$\ell_{dr} = \prod_t W_{rt}^{K_{dt}} e^{-W_{rt}} \tag{A.4}$$

Since each term in the product is small, there is a significant risk of underflow errors in the computation. To avoid this, the logarithm is calculated

$$\log(\ell_{dr}) = \sum_t [K_{dt} \log(W_{rt}) - W_{rt}] \tag{A.5}$$

An extremely useful property of this sum is that the first term is only non-zero when K_{dt} is non-zero. Also, the second term is independent of d, meaning it needs to be calculated only once (as `u[r]`). Thus, the calculation of $\log(\ell_{dr})$ has a time complexity which scales as the total number of photons in the data set, and not as the number of pixels times the number of frames (which is much larger for sparse data). This lowers the time required per iteration significantly. First `u[r]` is calculated,

for $r < R$ **do**

```
u[r] = - sub_total[r] * rescale
```

end for

To avoid underflows on exponentiation, one needs to exponentiate $[\log(\ell_{dr}) - m_d]$ where $m_d = \max_r \log(\ell_{dr})$. Thus, at least one orientation r will exponentiate to 1. Since, ℓ_{dr} is normalized over r, any r-independent factor cancels out. Here is the code to calculate $\log(\ell_{dr})$ and m_d. This is the most time-consuming computation.

```
for r < R do
    generate_rescaled_view(r, rescale)
    for d < D do
        log_prob[r][d] = u[r]
        for non-zero pixels t do
            log_prob[r][d] += count[t] * log_view[t]
        end for
        if log_prob[r][d] > max_prob[d] then
            max_prob[d] = log_prob[r][d]
        end if
    end for
end for
```

The next step is to get ℓ_{dr} by exponentiation and to calculate the normalization factor

```
for r < R do
    for d < D do
        prob[r][d] = exp(-log_prob[r][d] - max_prob[d])
        prob_sum[d] += prob[r][d]
    end for
end for
```

P_{dr} is used to calculate W'_{rt} using the formula

$$W'_{rt} = \frac{\sum_d P_{dr} K_{dt}}{\sum_d P_{dr}} \tag{A.6}$$

Again the numerator is calculated only over the non-zero pixels. These updated views are then merged to obtain the updated 3D model as described in the previous section. The outline of the process is as follows

for $r < R$ **do**
  for $d < D$ **do**
    `prob[r][d] /= prob_sum[d]`
    `sum_d += prob[r][d]`
    for non-zero pixels t **do**
      `updated_view[t] += count[t] * prob[r][d]`
    end for
  end for
  for $t < T$ **do**
    `view[t] /= sum_d`
  end for
  `merge_view(r)`
end for

A.2.3 Monitoring iterations

It is important to monitor the progress of the iterative process, both to catch errors as well as to compare across different data sets and reconstruction parameters. The primary tool used was the difference between one iteration and the next. This goes to zero as the iteration converges before bottoming out at machine precision. The slope of the curve just before convergence depends on the quality of the signal per frame. Another number calculated was the mutual information between r and d used to determine the "peakiness" of P_{dr}. The mutual information is defined as

$$\sum_{r,d} [P_{dr} \log(P_{dr}/P_r)] \tag{A.7}$$

where P_r is the probability of having a given orientation, taken to be uniform ($P_r = 1/R$). This can be understood to be the the entropy reduction of the rotational distribution, relative to the uniform distribution, and averaged over data frames. The higher the mutual information, the sharper the distribution of P_{dr}. This depended on just the number of photons/frame regardless of whether it was signal or background. So while convergence was slower for high background, the mutual information was higher.

A.3 Post-processing

After the model has converged, the peaks need to be integrated to generate a table of *hkl*-intensities. These would then be used to phase the structure. One can

also examine the reconstructions for quality in various ways. Another important requirement is the ability to compare two reconstruction for similarity. These are detailed below.

A.3.1 Peak integration

In the data set collected in the experiment of Chapter 4, the crystal had bcc symmetry and the *hkl* spacing was chosen to be 5 voxels. The peaks were integrated using a sphere around each lattice point (even the forbidden ones) of radius 2 voxels. This covered approximately 27% of the volume. A smaller radius would have been chosen if the peaks were sharper, but the polychromaticity of the beam broadened out the peaks.

A.3.2 Quality assessment

After the peaks have been integrated, they should be examined to see if they deviate from the known crystal symmetry. This is a useful tool to diagnose any misalignment issues, either in the detector panels or in the experimental geometry.

Another metric is the Bragg to diffuse scattering ratio as a function of q. This is the ratio of average intensity of a Bragg voxel to a non-Bragg voxel in a q-shell. Bragg voxels are defined to be those inside the integration volume. If this ratio is

high, most of the intensity is in Bragg peaks and that means that the orientations have been correctly identified. However, if this ratio is close to 1, the peaks are indistinguishable from the diffuse background. This essentially means that the algorithm could not find the peaks over the background. This could be either due to relatively high background, especially at high resolutions where the peaks are weak, or due to insufficient photons/frame leading to inability to recover orientations. The latter case would be indicated by the inability to resolve even the low-resolution peaks.

A.3.3 Intensity comparisons

The best way to compare the intensity tables (hkl - I) from two data sets is to determine the phases and compare the electron densities. If this is not possible, as in the experiment in Chapter 4, it becomes hard to compare just the intensities. Not all differences are equally important in influencing the electron density map, which is what is ultimately of interest. Thus, it may prove helpful to compare the Patterson maps. A Patterson map is the cyclic autocorrelation of the electronic density, which is obtained by inverse Fourier transforming the hkl - I table. These are much more amenable to visual comparisons and significant differences in two reconstructions show up clearly in the map.

BIBLIOGRAPHY

[1] Kartik Ayyer, Hugh T. Philipp, Mark W. Tate, Veit Elser, and Sol M. Gruner. Real-space x-ray tomographic reconstruction of randomly oriented objects with sparse data frames. *Opt. Express*, 22(3):2403–2413, Feb 2014.

[2] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 02 1970.

[3] Donald H Bilderback, Joel D Brock, Darren S Dale, Kenneth D Finkelstein, Mark A Pfeifer, and Sol M Gruner. Energy recovery linac (erl) coherent hard x-ray sources. *New Journal of Physics*, 12(3):035011, 2010.

[4] Sébastien Boutet, Lukas Lomb, Garth J Williams, Thomas RM Barends, Andrew Aquila, R Bruce Doak, Uwe Weierstall, Daniel P DePonte, Jan Steinbrener, Robert L Shoeman, et al. High-resolution protein structure determination by serial femtosecond crystallography. *Science*, 337(6092):362–364, 2012.

[5] Henry N Chapman, Petra Fromme, Anton Barty, Thomas A White, Richard A Kirian, Andrew Aquila, Mark S Hunter, Joachim Schulz, Daniel P DePonte, Uwe Weierstall, et al. Femtosecond x-ray protein nanocrystallography. *Nature*, 470(7332):73–77, 2011.

[6] R.R. Coifman, Y. Shkolnisky, F.J. Sigworth, and A. Singer. Graph laplacian tomography from unknown random projections. *Image Processing, IEEE Transactions on*, 17(10):1891–1899, Oct 2008.

[7] V. Elser. Noise limits on reconstructing diffraction signals from random tomographs. *Information Theory, IEEE Transactions on*, 55(10):4715–4722, Oct 2009.

[8] Dimitrios Giannakis, Peter Schwander, and Abbas Ourmazd. The symmetries of image formation by scattering. i. theoretical framework. *Opt. Express*, 20(12):12799–12826, Jun 2012.

[9] Katherine S Green, Hugh T Philipp, Mark W Tate, Joel T Weiss, and Sol M Gruner. Calibration and post-processing for photon-integrating pixel array detectors. *Journal of Physics: Conference Series*, 425(6):062009, 2013.

[10] James M. Holton. A beginner's guide to radiation damage. *Journal of Synchrotron Radiation*, 16(2):133–142, Mar 2009.

[11] G. Huldt, A. Szke, and J. Hajdu. Diffraction imaging of single particles and biomolecules. *Journal of Structural Biology*, 144(12):219 – 227, 2003. Analytical Methods and Software Tools for Macromolecular Microscopy.

[12] I Johnson, A Bergamaschi, H Billich, S Cartier, R Dinapoli, D Greiffenberg, M Guizar-Sicairos, B Henrich, J Jungmann, D Mezza, A Mozzanica, B Schmitt, X Shi, and G Tinti. Eiger: a single-photon counting x-ray detector. *Journal of Instrumentation*, 9(05):C05032, 2014.

[13] Lucas J. Koerner and Sol M. Gruner. X-ray analog pixel array detector for single synchrotron bunch time-resolved imaging. *Journal of Synchrotron Radiation*, 18(2):157–164, Mar 2011.

[14] N. D. Loh, M. J. Bogan, V. Elser, A. Barty, S. Boutet, S. Bajt, J. Hajdu, T. Ekeberg, F. R. N. C. Maia, J. Schulz, M. M. Seibert, B. Iwan, N. Timneanu, S. Marchesini, I. Schlichting, R. L. Shoeman, L. Lomb, M. Frank, M. Liang, and H. N. Chapman. Cryptotomography: Reconstructing 3d fourier intensities from randomly oriented single-shot diffraction patterns. *Phys. Rev. Lett.*, 104:225501, Jun 2010.

[15] Ne-Te Duane Loh and Veit Elser. Reconstruction algorithm for single-particle diffraction imaging experiments. *Phys. Rev. E*, 80:026705, Aug 2009.

[16] Richard Neutze, Remco Wouts, David van der Spoel, Edgar Weckert, and Janos Hajdu. Potential for biomolecular imaging with femtosecond x-ray pulses. *Nature*, 406(6797):752–757, 2000.

[17] Z Otwinowski and W Minor. Processing of x-ray diffraction data. *Methods enzymol*, 276:307–326, 1997.

[18] H.T. Philipp, L.J. Koerner, M.S. Hromalik, Mark W. Tate, and Sol M. Gruner. Femtosecond radiation experiment detector for x-ray free-electron laser (xfel) coherent x-ray imaging. *Nuclear Science, IEEE Transactions on*, 57(6):3795–3799, Dec 2010.

[19] Hugh T. Philipp, Kartik Ayyer, Mark W. Tate, Veit Elser, and Sol M. Gruner. Solving structure with sparse, randomly-oriented x-ray data. *Opt. Express*, 20(12):13129–13137, Jun 2012.

[20] Hugh T Philipp, Mark W Tate, and Sol M Gruner. Low-flux measurements with cornell's lcls integrating pixel array detector. *Journal of Instrumentation*, 6(11):C11006, 2011.

[21] Sjors HW Scheres, Haixiao Gao, Mikel Valle, Gabor T Herman, Paul PB Eggermont, Joachim Frank, and Jose-Maria Carazo. Disentangling conformational states of macromolecules in 3d-em through likelihood optimization. *Nature Methods*, 4(1):27–29, 2007.

[22] Claude Elwood Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.

[23] Francesco Stellato, Dominik Oberthür, Mengning Liang, Richard Bean, Cornelius Gati, Oleksandr Yefanov, Anton Barty, Anja Burkhardt, Pontus Fischer, Lorenzo Galli, Richard A. Kirian, Jan Meyer, Saravanan Panneerselvam, Chun Hong Yoon, Fedor Chervinskii, Emily Speller, Thomas A. White, Christian Betzel, Alke Meents, and Henry N. Chapman. Room-temperature macromolecular serial crystallography using synchrotron radiation. *IUCrJ*, 1(4), Jul 2014.

[24] M W Tate, D Chamberlain, K S Green, H T Philipp, P Purohit, C Strohman, and S M Gruner. A medium-format, mixed-mode pixel array detector for kilohertz x-ray imaging. *Journal of Physics: Conference Series*, 425(6):062004, 2013.

[25] Pierre Thibault, Martin Dierolf, Andreas Menzel, Oliver Bunk, Christian David, and Franz Pfeiffer. High-resolution scanning x-ray diffraction microscopy. *Science*, 321(5887):379–382, 2008.

[26] W. Vernon, M. Allin, R. Hamlin, T. Hontz, D. Nguyen, F. Augustine, S. M. Gruner, Ng. H. Xuong, D. R. Schuette, M. W. Tate, and L. J. Koerner. First results from the 128x128 pixel mixed-mode si x-ray detector chip, 2007.

[27] Eric W. Weisstein. *Secant Method*. From MathWorld–A Wolfram Web Resource.

[28] Thomas A. White, Richard A. Kirian, Andrew V. Martin, Andrew Aquila, Karol Nass, Anton Barty, and Henry N. Chapman. *CrystFEL*: a software suite for snapshot serial crystallography. *Journal of Applied Crystallography*, 45(2):335–341, Apr 2012.

[29] Jennifer L. Wierman, Jonathan S. Alden, Chae Un Kim, Paul L. McEuen, and Sol M. Gruner. Graphene as a protein crystal mounting material to reduce background scatter. *Journal of Applied Crystallography*, 46(5):1501–1507, Oct 2013.